ANDREIA FILIPA MOREIRA AGUIAR BRANDÃO

DEMOGRAPHY OF MODERN HUMANS IN SOUTHEAST ASIA (SEA): A GENETIC APPROACH

Tese de Candidatura ao grau de Doutor em Ciências Biomédicas submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, Porto, Portugal

Orientador - Doutora Luísa Pereira
Categoria - Investigadora
Afiliação - Instituto de Investigação e Inovação em Saúde, Universidade do Porto; Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Porto, Portugal

Coorientador - Doutor Pedro Soares
Categoria - Professor Assistente/ Investigador
Afiliação – Centro de Biologia Molecular e Ambiental, Escola de Ciências da Universidade do Minho, Braga, Portugal

Coorientador - Prof Martin Richards
Categoria - Research Chair
Afiliação - University of Huddersfield, Huddersfield, England

Cotutor - Professor Doutor Manuel Teixeira
Categoria - Professor Catedrático Convidado
Afiliação - Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, Porto, Portugal

O trabalho apresentado nesta tese foi realizado no Instituto de Patologia e Imunologia Molecular da Universidade do Porto (Ipatimup), no Instituto de Investigação e Inovação da Universidade do Porto (i3S) e na Universidade de Huddersfield.

Este trabalho foi financiado pela Fundação para a Ciência e Tecnologia (FCT) através de uma bolsa de doutoramento (SFRH/BD/78990/2011) e do projeto (PTDC/IVC-ANT/4917/2012).

Governo da República Portuguesa

UNIÃO EUROPEIA
Fundo Social Europeu

PRECEITOS LEGAIS

De acordo com o disposto no art.º 8º do Decreto-Lei n.º 388/70, nesta dissertação foram utilizados resultados dos trabalhos publicados ou em preparação abaixo indicados. No cumprimento do disposto no referido Decreto-Lei, a autora desta tese declara que interveio na conceção e na execução do trabalho experimental, na interpretação e discussão dos resultados e na redação dos manuscritos publicados, sob o nome Brandão A:

- *Soares P, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, Mormina M,* ***Brandão A,*** *Fraser RM, Wang T-Y, Loo J-H, Snell C, Ko T-M, Amorim A, Pala M, Macaulay V, Bulbeck D, Wilson JF, Gusmão L, Pereira L, Oppenheimer S, Lin M, Richards MB (2016) Resolving the ancestry of Austronesian-speaking populations.* Hum Genet, 135(3), 309-326. DOI: 10.1007/s00439-015-1620-z

- ***Brandão A,*** *Eng KK, Rito T, Cavadas B, Bulbeck D, Gandini F, Pala M, Mormina M, Hudson B, White J, Ko T-M, Saidin M, Zafarina Z, Oppenheimer S, Richards MB, Pereira L, Soares P. Quantifying the* legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia. Hum Genet (Epub ahead of print). DOI 10.1007/s00439-016-1640-3

- **Brandão A,** Cavadas B, Bulbeck D, Hudson B, White J, Chia S, Saidin M, Zafarina Z, Oppenheimer S, Pereira L, Richards MB, Soares P. The genome-wide landscape of Island Southeast Asia. *In preparation*

"*Two roads diverged in a wood, and I -*
I took the one less travelled by,
And that has made all the
difference."

Robert Frost,1920

AGRADECIMENTOS/ ACKNOWLEDGEMENTS

Ao longo destes quatro anos, muitas foram as pessoas que contribuíram para o meu crescimento pessoal e profissional. As experiências compartilhadas entre todas as pessoas envolvidas serão para sempre lembradas. A todos, o meu muito obrigada!

Gostaria de fazer um agradecimento especial aos meus supervisores Luísa Pereira e Pedro Soares por me terem dado a oportunidade de realizar este trabalho e por toda a disponibilidade e entusiasmo que sempre demonstraram. O vosso apoio fez-me crescer tanto a nível profissional, como pessoal.

I would also like to thank Martin Richards who kindly agreed to co-supervise my work, for his mentorship and continuous support. I was fortunate to have his guidance, and I am truly grateful for the many hours of advice over these years.

À Fundação para a Ciência e a Tecnologia por me conceder uma bolsa de doutoramento (SFRH/BD/78990/2011), sem a qual a realização deste trabalho não teria sido possível. Bem como, aos Professores Doutores Manuel Teixeira e Eduardo Rocha, ao Programa Doutoral em Ciências Biomédicas do Instituto de Ciências Biomédicas Abel Salazar, por me terem dado a oportunidade de seguir para Doutoramento.

A realização deste projeto não teria sido possível sem os meus colegas e amigos do IPATIMUP, e em particular do grupo Genetic Diversity, por quem nutro um carinho muito especial. Obrigado pela vossa ajuda e companheirismo. Um obrigada especial ao Bruno, por ter estado sempre presente quando precisei. Obrigada pela tua paciência e por me teres feito descobrir o mundo da bioinformática. Sem ti não teria sido capaz!

Às minhas meninas Andreia, Catarina, Joana, Marina, Marisa, Natália, Patrícia, Sílvia, Sofia e Susana, por todas as gargalhadas e momentos de amizade que partilhamos. Obrigada pelo vosso otimismo nos períodos mais tumultuosos, e por me fazerem ver sempre (ou quase sempre) o lado bom de tudo o que vivemos. Os nossos momentos, conversas e palhaçadas, são as melhores recordações que levo deste período da minha vida (L&L para sempre!).

I also owe a great deal of thanks to my colleagues from the Archaeogenetics Research Group at the University of Huddersfield. Everyone has provided me with extensive assistance over the course of this project, and they have my utmost gratitude

for their help. A special thanks to Maria Pala and Martin Carr, the conclusion of this project would not have been possible without their support and guidance.

Thanks to all collaborators and all of the co-authors of the publications presented in this thesis for their contributions.

A todos os meus amigos (vocês sabem quem são), por me aturarem e estarem sempre presentes, mesmo nas fases mais bipolares. Em especial, à Carla, minha alma gémea de sempre e para sempre, por tantos anos de amizade e carinho (já lá vão perto de 20 anos) e por tudo o que já passamos juntas.

À minha família, agradeço todo o amor e carinho incondicional ao longo de todos estes anos. Obrigada por me terem dado asas para poder voar e por me fazerem acreditar que nada é impossível, que basta sonhar e batalhar.

E por fim ao Daniel, tu foste o meu pilar nesta fase da minha vida. Sem ti, não teria conseguido manter a minha sanidade mental durante a escrita desta tese (assim como em tantos outros momentos do doutoramento!). Obrigada por me fazeres ver a vida de uma forma mais feliz e por seres, acima de tudo, o meu melhor amigo e o amor da minha vida.

Um sincero obrigado a todos,
Andreia

ABSTRACT

Understanding the global patterns of human genetic variation is of fundamental interest not only to evolutionary and anthropological sciences, but also in the biomedical research field. Despite the considerable amount of published studies, the genetic history of the populations of Southeast Asia (SEA), in particular of Island Southeast Asia (ISEA) is still not fully enlightened. For decades, the most widely accepted model for the peopling of ISEA involved a first colonization around 50,000-60,000 years ago, followed by a large population replacement in the mid-Holocene by Austronesian-speaking rice agriculturists from Taiwan. Recently, several studies have shown that this model does not completely account for the complex current population structure in ISEA, and have highlighted the dramatic postglacial climate changes in the Late Pleistocene/Early Holocene as the driving force for population movements in the region. Therefore, this study aimed to conduct a comprehensive genetic analysis, combining mitochondrial DNA (mtDNA) and genome-wide data, to test these two partially competing hypotheses and, ultimately, to provide a better resolved picture of population movements in the region.

The genetic analyses performed here revealed a clear signal of those two migratory events, though each one with different impact for the population gene pool. The extensive founder analyses of mtDNA control region revealed that the Late Pleistocene and Early Holocene were the key periods that shaped the mitochondrial diversity of ISEA, contributing to almost half of the lineages of the current gene pool. The results also showed that approximately 20% of present-day maternal lineages trace most likely to mid-Holocene Neolithic dispersals from both Mainland Southeast Asia (MSEA) and South China, via Taiwan. The mtDNA founder clusters associated with the possible migratory events in the region were further analysed at higher resolution through complete sequencing. The phylogenetic analyses of the whole-mtDNA sequences confirmed the picture initially obtained: lineages belonging to mtDNA haplogroups B4a1a, B5b1c, F3b1, B4c1b2a2, N9a6, R9c1a and E dispersed in the Late Pleistocene and Early Holocene, most likely due to massive migrations triggered by climate change and sea-level rises; while lineages B4b1a2, F1a4a, Y2a1, D5b1c1, M7c3c and M7b3 most likely spread in the mid-Holocene from Taiwan. Notably, the results also revealed that all lineages that showed an “out-of-Taiwan” ancestry in ISEA trace directly to South China, where the putative initial spread of rice-agriculturists across SEA took place. The genome-wide analysis confirmed that SEA populations have a highly complex genetic structure,

translated in several layers of genetic admixture, most probably as a result of thousands of years of extensive gene flow and adaptations to new environments. The autosomal data revealed a remarkably consistent picture with the modern mitogenome patterns, showing a distinct genetic composition between continental and insular SEA. Overall, most of MSEA populations, with the exception of the Malays, have a mixed ancestry between South and East Asian populations. However, in the case of ISEA populations, they presented a very interesting complex genetic structure, with two clear distinct ancestral genetic patterns: the western populations evidenced ancestral components related to MSEA, whereas the genetic structure of the eastern populations is more consistent with an influx of Taiwanese/Philippine and Papuan populations.

Altogether, this study offers not only new insights into the complex population genetic structure of Southeast Asia, but also new substantial genetic data for future population studies, namely for the evaluation of selection events and their contribution to the adaptation to differential environmental pressures across this geographical region.

RESUMO

A compreensão dos padrões globais da variação genética humana é de interesse fundamental não apenas em estudos evolutivos e antropológicos, mas também no campo da investigação biomédica. Apesar do número considerável de estudos prévios, a história genética das populações do Sudeste Asiático (SEA), em particular das populações das ilhas (ISEA), ainda não se encontra completamente esclarecida. Durante décadas, o modelo mais amplamente aceite para a colonização de ISEA envolvia uma primeira uma colonização por volta dos 50.000-60.000 anos atrás, seguida por uma extensa re-população em meados do Holoceno, por agricultores de arroz de língua Austronésia, oriundos de Taiwan. Recentemente, vários estudos têm demonstrado que este modelo não permite explicar totalmente a complexa estrutura populacional observada atualmente em ISEA, e têm apontado as dramáticas mudanças climáticas pós-glaciares no final do Pleistoceno/início do Holoceno como a força principal para os movimentos populacionais na região. Neste sentido, este estudo teve como objetivo efetuar uma extensiva análise genética, combinando dados mitocondriais (mtDNA) e autossómicos, de forma a testar estas duas hipóteses parcialmente concorrentes e, em última instância, fornecer uma visão integral dos movimentos populacionais que refletem a caracterização genética da região.

A análise realizada revelou claros sinais da ocorrência dos dois eventos migratórios referidos, embora o impacto de cada um para o *pool* genético da população seja diferente. A extensa análise fundadora da região controlo do mtDNA revelou que o Pleistoceno Superior e o ínicio do Holoceno foram os períodos-chave que moldaram a diversidade mitocondrial de ISEA, contribuindo para quase metade das linhagens do *pool* genético atual. Os resultados também mostraram que aproximadamente 20% das linhagens maternas atuais encontram-se associadas a dispersões em meados do Holoceno, oriundas tanto do Sudeste Asiático continental (MSEA) como do Sul da China, via Taiwan. Os *clusters* de fundadores mitocondriais que foram associados com os possíveis eventos migratórios na região foram analisados com maior resolução por sequenciação completa. As análises filogenéticas das sequências completas de mtDNA confirmaram o quadro obtido inicialmente: as linhagens pertencentes aos haplogrupos B4a1a, B5b1c, F3b1, B4c1b2a2, N9a6, R9c1a e E dispersaram-se no final do Pleistoceno e início do Holoceno, muito provavelmente devido a migrações em massa desencadeadas pelas alterações climáticas e pelo aumento do nível da água do mar;

enquanto as linhagens B4b1a2, F1a4a, Y2a1, D5b1c1, M7c3c e M7b3 dispersaram-se muito provavelmente a meio do Holoceno, vindas de Taiwan. Notavelmente, os resultados também revelaram que estas linhagens associadas com a chamada migração "out-of-Taiwan" apresentam ascendência direta no Sul da China, região onde hipoteticamente a dispersão dos agricultores de arroz pelo Sudeste Asiático se iniciou.

A análise da variação do genoma confirmou que as populações de SEA possuem uma estrutura genética complexa, traduzida em várias camadas de mistura, muito provavelmente como resultado de milhares de anos de extenso fluxo génico e adaptação a novos ambientes. Os dados autossómicos apresentaram uma imagem notavelmente consistente com os padrões mitocondriais modernos, revelando uma composição genética distinta entre as populações continentais e insulares de SEA. No geral, a maioria das populações de MSEA, com exceção dos Malaios, apresentaram uma ascendência partilhada entre populações do Sul e da Ásia Oriental. Já as populações de ISEA revelaram uma complexa estrutura genética muito interessante, com dois padrões genéticos ancestrais distintos: as populações ocidentais evidenciaram componentes ancestrais relacionadas com MSEA, ao passo que a estrutura genética das populações mais orientais é mais consistente com um influxo de populações Taiwanesas/Filipinas e Papuas.

De um modo geral, este estudo fornece não só novas perspetivas sobre a estrutura genética populacional do Sudeste Asiático, mas também novos dados genéticos substanciais para futuros estudos populacionais, nomeadamente na avaliação de eventos de seleção e no seu contributo para a adaptação a diferentes pressões ambientais nesta região geográfica.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

CHAPTER 3

Paper I - Resolving the ancestry of Austronesian-speaking populations

Paper II - Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia.

Paper III - The genome-wide landscape of Island Southeast Asia

APPENDICES

Appendix A - Supplementary Information of Paper I

Appendix B - Supplementary Information of Paper II

Appendix C - Supplementary Information of Paper III

LIST OF TABLES

CHAPTER 3

APPENDICES

Appendix B - Supplementary Information of Paper II

Appendix C - Supplementary Information of Paper III

LIST OF ABBREVIATION

A – Adenine

Bp – Base pair

C – Cytosine

Chr – Chromosome

CRS – Cambridge Reference Sequence

DNA – Deoxyribonucleic acid

ENCODE – Encyclopedia of DNA Elements

G – Guanine

Gb - Gigabases

GWAS – Genome-wide association studies

HGDP – Human Genome Diversity Project

HKY85 – Hasegawa-Kishino-Yano 85

HVR-I – Hypervariable region I

HVR-II – Hypervariable region II

ISEA – Island Southeast Asia

JC69 – Jukes–Cantor 69

K80 – Kimura 80

Ka – Thousands years ago

Kb - Kilo bases

LD – Linkage disequilibrium

LGM – Last Glacial Maximum

MCMC – Markov chain Monte Carlo

ME – Minimum evolution

MJ – Median-joining

ML – Maximum Likelihood

MP – Maximum Parsimony

MRCA – Most recent common ancestor

mRNA – Messenger RNA

MSEA – Mainland Southeast Asia

MSY – Male Specific region

mtDNA – mitochondrial DNA

nDNA – Nuclear DNA

Ne – Effective population size

NJ – Neighbor Joining

NMTCN – Nusantao Maritime Trading and Communication Network

NRY – Non-Recombining Region of the Y chromosome

OOT – Out-of-Taiwan

PAR – Pseudoautosomal regions

PCA – Principal components analysis

rCRS – revised Cambridge Reference Sequence

RM – Reduced-median

rRNA – Ribosomal RNAs

SEA – Southeast Asia

SNPs – Single Nucleotide Polymorphisms

STRs – Short tandem-repeat polymorphisms

T – Thymine

TMRCA – Time More Recent Common Ancestor

TN93 – Tamura-Nei 93

t-RNA – Transfer-RNA

U – Uracil

UPGMA – Unweighted pairwise group of multiple alignments

YCC – Y Chromosome Consortium

CHAPTER 1 - GENERAL INTRODUCTION

1. RECONSTRUCTION OF THE HUMAN POPULATION HISTORY

Over the last decades many studies started to focus on the discovery and characterization of human genetic variation, due to its importance for reconstructing the human evolutionary history and for understanding the genetic basis of human diseases. The pattern of the human genetic variation is modeled by historical demographic events (including population size changes, admixture events, and migrations) and by genetic factors, such as mutation and recombination rates. Thus the study of the different patterns of human genetic diversity allows us to infer not only how human adaptation to new environments shaped variation in the genome, but as well its implications for the fields of functional genomics, pharmaceutical applications and medicine development.

The advances in DNA sequencing techniques and the implementation of large-data research tools led to significant developments in population genetic studies. An example is the research being conducted in Southeast Asia (SEA), where several studies have been contributing invaluable information for a better understanding of its complex prehistory. The majority of these studies only focused on either maternal (mitochondrial DNA; mtDNA) or paternal lineages (Y chromosome), but a few are beginning to provide high-resolution screenings of specific major lineages as well as dealing with genome-wide information. Nevertheless, the huge diversity of this geographic region calls for a much intensive systematic analysis at a high-resolution level. The aim of this work is to shed light on the SEA human genetic history, by performing a highly informative phylogeographic analysis of a large number of maternal lineages and by analyzing the genome-wide diversity through the characterization of a chip containing 700,000 polymorphisms.

1.1. THE STUDY OF HUMAN EVOLUTION

Since early times researches have been attempting to comprehend the major questions regarding the origins and evolution of our species. Several fields have emerged to seek this quest, in particular, archaeology, climatology, linguistics, and anthropology. Taking advantage of their diverse scopes and research methods, the combination of

results gathered by those fields allows to obtain a comprehensive picture of the human population evolution [1]. More recently, a new field has emerged focusing the study of the population genetic variation. Genetic information provides a distinct way of viewing human evolution, since the modern human genome contains imprinted the effect of various evolutionary forces (expansions, bottlenecks and population migrations), in the form of altered gene frequencies which are transmitted along successive generations [2].

1.1.1. Basic concepts of the genetic information

1.1.1.1. DNA molecule

The genetic information encoded in the nucleic acids, is essential for cellular life, and it is transmitted to the daughter cells during cell division. The Deoxyribonucleic Acid (DNA) is a double helix polymer constituted by two polynucleotide strands. The nucleotide unit comprises a nucleobase, a sugar, and a phosphate. The backbone of the DNA strand consists in a phosphate linked by phosphodiester bond to a sugar, known as 2-deoxyribose, to which a base is attached (fig. 1A). The four bases found in DNA are classified as purines, double-ringed molecules called adenine (A) and guanine (G) or pyrimidines, single-ringed molecules named thymine (T) and cytosine (C). There is a third pyrimidine, called uracil (U), that substitutes thymine in Ribonucleic Acid or RNA molecules, and differs from thymine by lacking a methyl group on its ring .

DNA is composed by ‘anti-parallel’ strands, which means that one strand is in the 5′□to 3′□orientation and the other strand lines up in the 3′□to 5′□direction relative to the first strand. The two strands are linked together through a process known as hybridization, where the individual nucleotide pairs up with its ‘complementary base’, thus the adenine on one chain is always paired with thymine on the other chain and, likewise, guanine is always paired with cytosine (fig. 1B) [3, 4]. During cell division, the synthesis of new DNA molecules is called replication. In this process the two DNA strands separate and each one of them serves as template for the production of complementary DNA strands, based on the pairing of the complementary bases. In this way, two identical DNA molecules, are formed for each daughter cell [1].

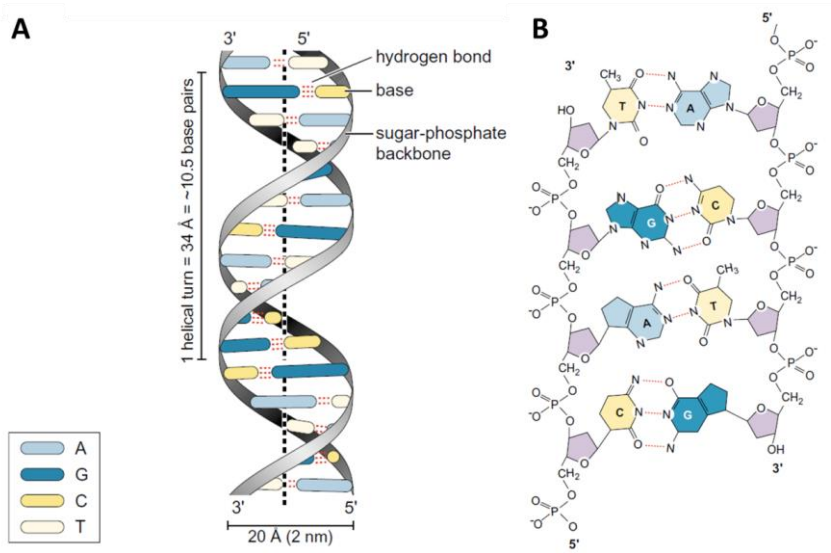


Figure 1. The structure of the DNA molecule. (A) Schematic representation of the DNA double helix. (B) Detailed structure of the DNA strands showing base pairing between pyrimidines (in yellow) and purines (in blue), and the phosphodiester linkages of the backbone. Adapted from [4].

1.1.1.2. Human genome organization

The human genome is composed of nuclear DNA (nDNA) and mtDNA. The nDNA is present in the nucleus of the cell and is organized into chromosomes, which are dense packets of DNA embracing protection proteins called histones. The nuclear human genome has approximately 3.2 gigabases (Gb) and consists of 23 pairs of chromosomes - 22 matched pairs of autosomal chromosomes and two sex determining chromosomes [1]. The sex chromosomes are either X,Y in males or X,X in females. On the other hand, the mtDNA is a small, multi-copy circular genome with only ~16.6 kb (kilo bases) located in cell's mitochondria [5].

The genetic material in the chromosomes comprises a set of non-coding intergenic regions and coding regions. The intergenic regions, mostly very highly repetitive sequences, were in the past thought to be transcriptionally silent DNA and, thus, of no value. However, recent studies showed that these non-coding regions contain functionally important elements, such as promoters and enhancers. In fact, genome-wide based projects (for example the ENCODE, which will be discussed in more detail further in this

chapter) revealed that these regions may function not solely as DNA elements, but they can also be transcribed into non-coding RNAs, possibly with regulatory functions [6-9]. Additionally, contrary to the past accepted notion that non-coding DNA evolves largely free from natural selection, some non-coding regions are highly conserved, suggesting that they are under strong selective constraints (negative selection) [10-12].

The coding regions, known as genes, are composed of introns, usually non-coding but containing sites for control of splicing (the process responsible for the removal of introns from the mature mRNA), exons, which are the coding regions for specific amino acid sequences, and also regulatory sequences that regulate the expression of the gene. The genes are transcribed into complementary pieces of messenger RNA (mRNA); in this case only one DNA strand (sense strand) acts as template [4]. The mRNA migrates from the nucleus to the cytoplasm, where it is translated into protein according to a code of triplets of the DNA bases (C, A, T, G), called codons.

1.1.2. Factors controlling genetic diversity

The genetic diversity of a population, i.e., the frequency of the polymorphic genetic variants in a population, is affected by both natural selection and stochastic factors, such as mutations, recombination, migrations and genetic drift (random effects on allele frequencies). All these evolutionary forces affect both the genetic variation within populations and the genetic divergence between populations. Mutations, recombination and migrations (gene flow) may increase the genetic variation within populations, because they introduce new alleles to the population. Whereas, genetic drift may cause loss of genetic diversity, through fixation of alleles. Natural selection may either increase or decrease genetic diversity within populations.

1.1.2.1. Mutation

Genetic diversity arises primarily with changes in the nucleotide sequence. All changes producing new alleles are by definition considered mutations. These can result from spontaneous errors in normal cellular mechanisms that regulate chromosome

segregation, recombination and DNA replication or from covalent changes in structure due to the action of environmental chemical or physical agents.

Mutations can occur at a broad variety of levels, from single nucleotide polymorphisms (SNPs), through differences in sequence copy numbers, to large-scale changes in the structures of chromosomes, each at different rates and with different phenotypic consequences. At the smaller level, the different polymorphism throughout the human genome can be classified as sequences polymorphisms or length polymorphisms. The first, resulted from changes of one or more bases in a DNA sequence, while the second, resulted from insertions or deletions of one or more nucleotides. The length polymorphisms are more frequently observed in repetitive DNA, like microsatellite DNA, also widely known as short tandem repeats (STRs) [1, 13]. STRs are tandem arrays of repeat units of 1-7 bp in length, and those that have a useful degree of polymorphism have a typical copy number of 10-30 [1].

1.1.2.1.1. Types of single nucleotide substitutions

SNPs are the most common contributors for the genetic variation in the human genome. SNPs can be classified into transitions and transversions based on the type of nucleotide substitution. The first comprises the substitution between pyrimidine nucleotides (e.g., C for T), or purine nucleotides (e.g., A for G), while the term transversion refers to the alteration of a purine for a pyrimidine and vice-versa. Single nucleotide indels, the gain or loss of a nucleotide, are also considered SNPs [5, 14].

The DNA substitutions in coding-regions can also be classified according to their effect on the protein. If a mutation changes the triplet codon into another that specifies the same amino acid, it is called a synonymous or silent substitution (though they are not always silent [15]). On the other hand, if it leads to the alteration of the amino-acid (usually substitutions at the second position and some at the first position of the codon) it is considered a non-synonymous mutation, and so, it can be subjected to natural selection [16]. The non-synonymous mutations are further classified into missense mutations, when it changes the codon to one that codifies a different amino acid from the previously encoded, and nonsense mutations, when it changes the codon into a termination codon (stop), leading to the premature ending of the translation process [15]. The indels occurring in exons also can change the protein: when in multiples of three (in-frame), by

the loss or gain of aminoacids; and, when not in multiples of three nucleotides, the reading frame is dramatically changed and the resulting protein is completely different.

1.1.2.1.2. The rate of mutations

The rate of mutations is not uniform throughout the genome. For instances, coding or regulatory regions in the genome are usually less polymorphic than the non-coding ones, because of their biological function, and subsequent natural selection pressure. The knowledge of the rates of the different types of mutations (described above) is essential in evolutionary genetic studies; in the sense that, these rates can be incorporated into the evolution models used to interpret patterns of diversity within and between species, and be used as molecular clock (further discussed in 1.1.3) [1, 17].

Methods for estimating mutation rates are preferentially based on a direct comparison of non-functional DNA sequences [1, 18], as for neutral mutations (with no effect on fitness), the mutation rate is expected to be equal to the rate of evolution. In that sense, the direct comparison of DNA sequences between species whose divergence times are known, can give an estimate of the mutation rate [1, 17]. The development of high-throughput DNA sequencing technologies have revealed that base substitutions, occurring at an average rate of the order of 10^{-8} per base per generation, are ~10 times more frequent than small indels [1]. Furthermore, there is a pronounced difference in the rate of base substitution between transitions and transversions, being the first almost three times more frequent than transversions (8.15^{-9} and 3.87^{-9} respectively [19]) [1]. STRs are fast evolving markers, with mutation rates about 10^{-3} per year [1, 20, 21].

1.1.2.2. Recombination

Recombination is one of the key evolutionary processes shaping the human genetic diversity. Recombination involves the reciprocal exchange (or crossover) of genetic material between two homologous chromosomes in meiosis, during a process called synapsis [22]. The exchange of genetic information between paired chromosomes inherited from each parent, generates novel combinations of the existing alleles in the same DNA molecule, known as haplotypes. Therefore, recombination breaks the

haplotype present in one generation to yield a new haplotype in the next, increasing the haplotype diversity, which is extremely important to gain evolutionary advantageous combinations of alleles and to eliminate deleterious mutations in populations [1].

Recombination is becoming increasingly used in population genetic studies. Some SNPs, especially if they are close together in the genome, have higher probability of being transmitted linked, since the occurrence of recombination between them will be lower. This non-random correlation of alleles at two or more loci is known as linkage disequilibrium (LD). The study of the decay of LD (which will be further discussed in this chapter), is becoming a promising tool in human population genetic studies, allowing to map candidate genes/SNPs (i.e., genetic mapping) and to estimate time of admixture between parental populations [1, 23, 24].

1.1.2.3. Migration

Another important factor that may increase the genomic variation in populations, by changing allele frequencies trough time, is the influx of genes from other populations, commonly called migration or gene flow (the outcome of a migrant contributing to the next generation in their new location) [1]. Sometimes, individuals from one population (source population) migrate and settle elsewhere (sink population), introducing previously non-existing genetic variants in the receiving population, or increasing their frequency.

Therefore, overall the migration has two major effects. On one hand, new alleles may arise in different populations due to rare mutational events, and these alleles can be spread to new populations by migration, thus increasing the genetic variation within the recipient population. And in another hand, migration prevents genetic divergence between different populations, because it tends to keep populations homogeneous in their allele frequencies, making the populations gene pools more similar [22].

The magnitude of change due to migration depends on both the extent and demographic/cultural aspects of migration (namely, the gender of the migrants) and the difference in the allele frequencies between the source and the sink populations. For instances, different marital residence patterns in the populations (whether they are patrilocal - women move to the husband's location; or matrilocal - the men move to the wife's location), generates different patterns of diversity between mtDNA (female lineages) and Y chromosomes (male lineages) [1].

1.1.2.4. Genetic drift

Genetic drift refers to changes in allele frequency that result from the random sampling of gametes from generation to generation in a population [2, 25]. The extent of change in allele frequency due to genetic drift is directly related to the effective population size (Ne, the equivalent number of those individuals who genetically contribute to the next generation in a population) [26]. In this sense, the chance effects are small in large populations, in comparison to the effects of other factors, such as gene flow and selection. Whereas, genetic drift can significantly affect the composition of small populations' gene pool for multiple generations.

Genetic drift may arise, simply because a population is kept small for a relatively long period of time due to limitations in food, space, or some other critical resource. A second way that genetic drift arises is through bottleneck, which is when a population undergoes a drastic reduction of effective size, and only few individuals survive (fig. 2) [26]. The small effective population size may also contribute to genetic drift trough a founder effect, consisting in the establishment of a population by only a small number of founder individuals. In that case, even if the population increases drastically, the gene pool diversity is limited to the one originally present in the founders (assuming no mutation or migration).

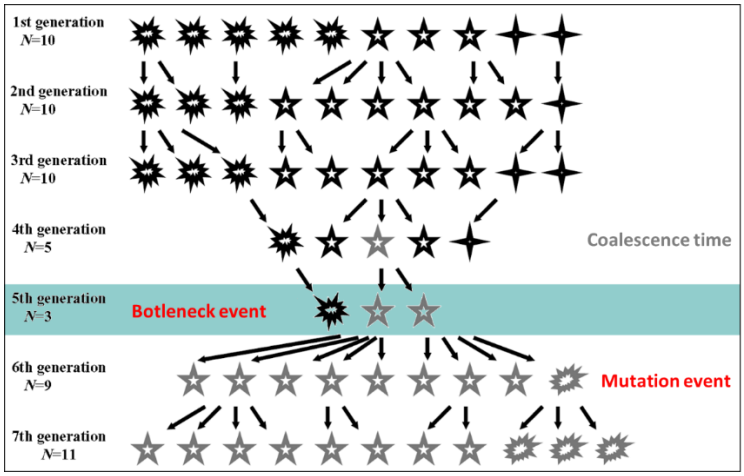


Figure 2. The basic principle behind population dynamics of alleles. Each symbol represents a different allele in a population. The figure exemplifies the fixation and loss of alleles during a population bottleneck event, and the grey symbols trace back to a single common ancestor, demonstrating the coalescence theory (to be introduced in section 1.1.3). N represents the number of individuals in the population. Adapted from [26].

Through random change, the allele frequency can undergo large fluctuations in a population, in an apparently unpredictable pattern, which may lead to new allelic variants to be either fixed in (reaching an allele frequency of 1.0, more rarely) or lost from (reaching an allele frequency of 0.0, more often) a population over time. This loss of genetic variation can only be restored by mutation or migration from another population.

1.1.2.5. Modelling pre-existing genetic variation by natural selection

Natural selection refers to the differential reproduction of individuals from different genotypes as a consequence of their different phenotypic characteristics. These fitness-enhancing traits (that improve an individual's chance of survival or reproductive success), are more likely to be passed on to the next generations, and therefore increase in prevalence in the population over time [1].

Natural selection may act in many different forms and with different intensities. The most common type of natural selection is the directional selection, in which an allele is favoured and so propagated (positive selection) or disfavoured and eliminated (negative selection, also called purifying selection) [27, 28]. The removal of deleterious mutations can also result in the occasional removal of neutral linked variation, a form of negative selection referred to as background selection. Genomic stretches under strong background selection, tend to be more conserved, with few, if any, non-synonymous mutations tolerated [28]. On the other hand, negative or purifying selection tend to be weaker for synonymous mutations with only mildly deleterious effects, allowing those mutations to accumulate at population level and to be maintained at a low frequency [29].

Positive selection (also known as Darwinian selection) promotes the emergence of new phenotypes, as the new mutations are advantageous in a given environment, increasing the fitness of the individuals [28-30].

Multiple alleles may be maintained at a given locus if they are advantageous individually or together, due to the effects of balancing selection. This type of selection, favouring the maintenance of genetic diversity in a population, might happen because of heterozygote advantage (i.e., overdominance) or frequency-dependent selection [31]. Additionally, if the two alleles that are being maintained conduct to opposing phenotypes in the population instead of an intermediate phenotypic effect, these alleles are under diversifying or disruptive selection. Contrastingly, if the intermediate phenotypes are the

ones favoured, whether by balancing selection of codominant alleles or by positive selection of alleles that underlie intermediate phenotypes, the selective regime is known as stabilizing selection [27].

1.1.3. The neutral theory of molecular evolution and the molecular clock

In 1968, Motoo Kimura formulated the neutral theory of molecular evolution [32], which claims that the majority of evolutionary changes at molecular level result from the random genetic drift of selectively neutral or nearly neutral mutations, rather than from natural selection [32, 33]. According to this theory, the majority of polymorphisms in the populations have little or no effect on fitness, so their fate (whether they are fixated or eliminated) is mostly determined by genetic drift. This theory does not reject the role of natural selection in cases of adaptive evolution, but assumes that only a small fraction of the molecular changes are adaptive. Furthermore, it predicts that in those few cases of selection, the prevailing mode is negative selection, instead of positive or balancing selection [1].

One of the most important outcomes of the neutral theory was the development of the concept of a molecular clock. The molecular clock was proposed by Emile Zuckerandl and Linus Pauling, in 1965, on the basis that the mutation rate for neutral mutations is relatively constant and equal among all organisms at all time, and therefore the rate of evolution is also approximately constant over all evolutionary lineages [34]. Accordingly, assuming that the degree of difference between the DNA sequences increases linearly with the time of divergence, if an absolute geological age (from archaeological evidences) is used as a calibration point of particular lineage divergence, then the mutation rate can be calculated, and, subsequently, all splits between different lineages traced on the tree can be dated [1, 34-36].

The molecular clock has become an important tool in evolutionary biology to reconstruct population history, however several studies have shown that molecular clocks may vary in different organisms and, even within a single organism between different genetic systems or lineages. For instances, the molecular clock for mitochondrial DNA is

faster than that for the nuclear genome, and we will explore this property more extensively later [22].

1.2. METHODS EMPLOYED TO STUDY HUMAN GENETIC DIVERSITY

The current human genetic diversity mirrors the history of past human population movements and adaptation to new environments. By characterising extant human genetic diversity we may infer the past of the human species. There are several methods that aim to characterise genetic diversity.

1.2.1. Inferring genetic distances

By definition, genetic distances are statistics that measure the similarity between sequences/individuals/populations, providing information on their relatedness [1, 22].

There are several methods to evaluate the genetic distance between two sequences or two populations (known as a pairwise distance). One of the most commonly used classical measure of genetic distance is F_{ST}. Working independently, Sewall Wright and Gustave Malécot introduced, in the late 40s and 50s, the *F* statistics as a tool for describing the partitioning of genetic diversity within and among populations (population structure) [37]. Population structure can be caused by the non-random mating of individuals, for instances, in small or isolated subpopulations. This isolation, eventually, leads to the partial genetic differentiation in subpopulations, as they undergo different evolutionary forces (e.g. genetic drift) [1]. Therefore, the level of population structure is highly dependent of the size of the metapopulation (the one that includes the partially differentiated subpopulations); large populations among which there is extensive gene flow tend to show little differentiation, whereas small populations among which there is little gene flow tend to be highly differentiated [37].

F_{ST} allow us to explore population structure by measuring the genetic differentiation among subpopulations. It is directly related to the total variance in allele

frequencies that occurs between subpopulations; in other words, it is directly related to the degree of resemblance among individuals within subpopulations relative to the entire population. This means that, if F_{ST} is small, then the allele frequencies within each population are very similar, most probably due to large amounts of gene flow between subpopulations; whereas if it is large, then allele frequencies are very different, which means that the subpopulations are highly differentiated [1, 37].

1.2.2. Phylogenetic inferences

The field of molecular phylogenetics has become a powerful tool for population genetics studies. The main objective of phylogenetics is to establish the evolutionary relationship between species, organisms or genes, by constructing phylogenies or trees [26, 38]. Basically, phylogenetic methods compare the similarities between different genetic lineages, in order to reconstruct the evolutionary relationships between them, which is graphically represented as a branching diagram, or tree, with branches joined by nodes. Each branch represents a genetic lineage through time, and each node represents the beginning of new lineages [35]. In particular, in a population genetics context, the nodes represent the birth events of individuals who are ancestral to the sampled population. The phylogenetic trees representing the evolutionary history of genes are known as gene trees, whereas if the tree establishes the evolutionary relationship among species, it is designated as species tree. It is important to emphasize that gene trees do not necessarily show the same structure as the species tree, because evolutionary forces reflects differently in both trees.

1.2.2.1. Key features of phylogenetic trees

The phylogenetic trees can be unrooted or rooted (fig. 3). An unrooted phylogenetic tree only reveals the relationships between lineages, it does not identify the most recent common ancestor (MRCA), or the direction of the evolutionary process. In opposition, a rooted tree shows the evolutionary relationship between a group of related lineages over time, based on the identification of the MRCA. The tree can be rooted by

two different methods: i) it can be assumed that the root lies at the midpoint of the longest branch on the tree (midpoint rooting), or ii) it can be added a distantly related taxa (outgroup) [26, 38]. For example, the phylogenetic tree of humans is usually rooted by the addition of chimpanzee lineages.

It is possible to establish three distinct types of relationship in a phylogenetic tree: i) if all lineages have a single evolutionary origin, they are called a monophyletic group; ii) if a group of lineages share the common ancestor but does not include all descendants of that common ancestor, it is considered a paraphyletic group; and iii) if the characters that support the phylogenetic group are not present in the MRCA, the group resulted from convergent evolution, and therefore they are designated as polyphyletic groups [38, 39].

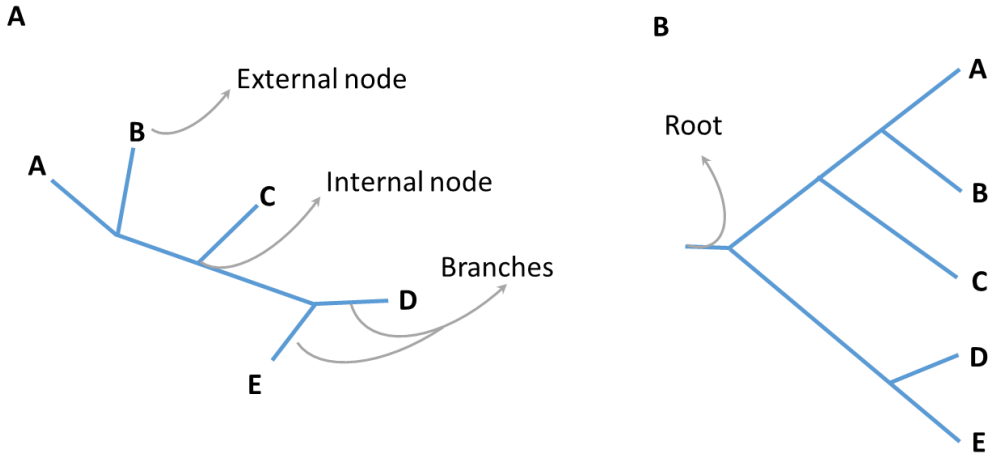


Figure 3. Schematic representation of random phylogenetic trees. (A) An unrooted tree with five external nodes. (B) Hypothetical rooted tree that can be drawn from the unrooted tree shown in A.

1.2.2.2. Phylogenetic tree reconstruction

There are several methods to infer phylogenetic relatedness, the main distinction between them being correlated with the way in which the genetic data is converted into numerical data to be mathematically analysed. The main methods in phylogenetic reconstruction can be classified as distance or character-based methods.

1.2.2.2.1. Distance-based methods

Distance-based methods employ an algorithm based on an evolutionary model, that converts sequence information into a distance matrix based on the number of differences between each pair of sequences in the dataset. The distance matrix is used to calculate the branch lengths connecting the sequences in the reconstructed phylogenetic tree. Distance based methods include unweighted pair group method using arithmetic averages (UPGMA) [40], neighbor joining (NJ) [41] and minimum evolution (ME)[42].

The UPGMA is probably the oldest and simplest method used for reconstructing phylogenetic trees from distance data, and was originally developed for phonetics. It reflects phenotypic similarities rather than evolutionary distances, therefore it assigns equal weight on the distance and assumes a randomized molecular clock. UPGMA is relatively a simple and fast method, however it behaves poorly if the assumption of an approximately constant rate of evolution among the lineages does not hold [26].

The NJ method is comparatively rapid and, generally, gives better results than the UPGMA method. NJ is the distance-based method most commonly used. It follows the minimum evolution concept, which means that it uses the least number of mutations required to obtain a given tree. The NJ starts by assuming that there is only one internal node from which all branches radiate in a star-like pattern, and calculates the length of the resulting tree. The algorithm sequentially examines all the possible pairs of neighbours, identifying the combination that yields the shortest tree. This process is repeated, introducing the shortest possible internal branches until the last pair is located, and the phylogenetic tree is reconstructed [26]. In the case of ME, as previously said, it seeks the tree with the minimum sum of branch lengths. ME fixates the internal nodes, by using the distance to external nodes, and then optimizes the internal branch lengths. A drawback of the ME method is that, in principle, all different tree topologies have to be investigated to find the minimum tree. Thus, although ME is more accurate (particularly for longer sequences) than the others distance-based methods, it is more time consuming than NJ, which generally yields very similar results .

Distance-based methods allow to rapidly and easily analyse relatively large datasets, however they also have significant disadvantages. For instance, NJ only produces a single phylogenetic tree, which may lead to loss of information [43].

1.2.2.2.2. Character-based methods

Character-based methods infer the most probable tree based on characters at each position in the sequence alignment. Maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference methods constitute examples of character-based methods. In these methods, each sequence position is considered one at a time to obtain a score and these scores are computed to identify the tree that either represents the minimum number of changes for maximum parsimony, the likelihood value for maximum likelihood and the posterior probability for Bayesian analysis [43]. All of them imply a model of evolution.

1.2.2.2.2.1. Maximum parsimony (MP) and Networks

MP method reconstructs the phylogenetic trees based on the minimum number of character changes along its branches required to explain the observed states at the terminal nodes. The main advantage of MP method is also its main disadvantage, since the number of changes on each branch is used as diagnosable units for each clade and branch lengths. In fact, serious biases can be introduced when mutational rates differ between conserved and hypervariable regions and/or the evolutionary rates are highly variable among different lineages leading to the overestimating of the evolutionary relationship. In these cases, maximum likelihood methods are best to infer reliable phylogenies [43].

Another issue when reconstructing parsimonious trees in human DNA phylogenetic analysis is the intraspecific short distance between individuals. Homoplasies (similar characters produced by convergent evolution) due to the occurrence of reversions (a character changes back to its previous state) or parallelisms (the same change occurred independently in two different taxa or groups of taxa), may create incompatibilities in the classical phylogenetic trees. In these situations, equally likely tree topologies, representing equally likely pathways of evolution, can be drawn from the same dataset. These loops within phylogenies (known as reticulations or cycles) can be represented in networks.

Similarly to tree phylogenetic reconstruction, networks can be inferred by different methods, namely, reduced-median (RM) and median-joining (MJ) networks. RM networks

resolves the reticulations by removing the network's least likely links from all the generated possible trees. The algorithm partitions the groups of haplotypes character-by-character, employing parsimony and frequency-based criteria, allied to the knowledge of site relative mutation rate (relative weights to the mutations, based on their occurrence rates [44]). This approach is usually applied for small sample sizes, and when homoplasies are frequent (therefore often applied in human mtDNA variation studies). Since this type of network contains all the equally likely trees, it can assist in identifying sequencing artefacts, which manifest themselves in incompatibilities in network substructures [45]. The MJ network algorithm is less accurate, but it can handle larger sets of genetic data, as well as multi-state markers such as amino acid sequences [46]. It begins by selectively combining the minimum spanning trees into a single (reticulate) network. Few consensus sequences (median vectors) of three mutually close sequences at a time can be added to attain the most parsimonious [46].

1.2.2.2.2.2. Maximum likelihood

Maximum likelihood methods examine different tree topologies and select the tree that maximizes the probability that a dataset fits the tree derived from that dataset, under a specified model of evolution. The ML algorithms search for the combination of branch lengths and evolutionary parameters that yields the highest likelihood score [39]. The ML inference is based on robust statistical foundations, which make this method a powerful tool for phylogenetic reconstruction. However, ML approach can be computationally very demanding, and therefore, it is limited to a relatively small dataset [26, 43].

1.2.2.2.2.3. Bayesian inference

Bayesian inference is also a likelihood method. But unlike ML, a further set of parameters (called priors) are input into the original model. The Bayesian inference produces a posterior probability distribution given the model of evolution, the observed characters in the dataset, and the specification of a set of assumptions (the prior probability distribution). All inferences concerning the parameters are then based on the estimation of the posterior distribution, mostly, using Markov Chain Monte Carlo

algorithms (MCMC algorithms). The MCMC is a computer simulation that generates a sample from a target distribution [35]. The Bayesian probability method is also computer intensive in constructing the phylogenies.

1.2.2.3. Models of evolution

Several substitution models of DNA sequence evolution have been proposed. These models assume that new mutations in the DNA sequences are completely independent from those previously existing [47], but differ in terms of the assumptions related with the rates of the nucleotide substitutions during evolution. The models of evolution are frequently used in molecular phylogenetic analyses, in particular in the Bayesian and maximum likelihood approaches to tree estimation.

The most commonly used are the JC69 (Jukes–Cantor, 69) [48], the K80 (Kimura, 80) [49] and the HKY85 (Hasegawa-Kishino-Yano, 85) [50] models. Both the JC69 and the K80 models predict equal frequencies of the four nucleotides, but the first assumes an equal substitution rate between any two nucleotides (whether pyrimidines or purines), whereas the K80 model accounts for the difference between the mutation rate for transitions (both purines/ pyrimidines) and transversions (from a purine to pyrimidine and vice-versa). The HKY85 model assumes that different sequence positions may evolve at different rates, in a relaxed nucleotide frequency (fig. 4).

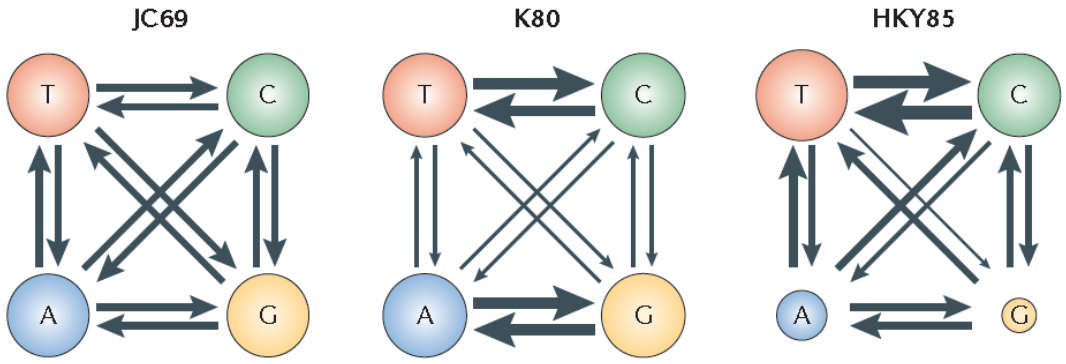


Figure 4. Substitution models. The figure illustrates the differences in the substitution rate between the four nucleotides (arrow thickness) and the nucleotide frequency (size of the circles) in the three models: JC69; K80 and HKY85. From [35].

There is a more complex model, the TN93 (Tamura-Nei, 93) model [51], which accounts for the difference between transitions and transversions, and differentiates the substitution among the different types of transition, i.e. purine and pyrimidine transitions [26, 35]. The most general neutral, independent and time-reversible substitution model for phylogenetic inference is the generalised time-reversible (GTR or also known as REV) model, firstly described by Simon Tavaré in 1986 [52] The GTR model is the general time reversible model that allows all substitutions (e.g., the four nucleotides A, C, G, and T) to occur at different rates, assuming that these relative rates remain constant across the tree, and that the root base composition is in equilibrium [53].

1.2.2.4. Dating TMRCA

The basic concept underlying the coalescent process is that, in the absence of selection, random lineages, at some point in history, trace back to a single ancestor, the MRCA. The MRCA represents the root of a phylogenetic tree, and therefore the estimation of the time to the root of a phylogeny, translates into the estimation of time of the MRCA (TMRCA) [26, 54].

There are several methods to estimate the TMRCA, one simple method is to calculate the average number of mutations between a set of individuals and a specified common ancestor (a statistic method referred as rho (ρ)) [1, 55]. This is a relatively straightforward statistical approach, that does not account for the prehistoric demography and population structure affecting the molecular evolution, therefore it is called a “model-free” statistical approach [47, 56]. It is important to stress out that since the ρ statistic requires an haplotype phylogeny, the correct specification of the phylogenetic topology is essential [47]. Another key feature in dating the MRCA is the generation time, which is essential to convert the date estimate from generations into years. The most commonly used generation time estimates used for modern humans tend to vary between 20 and 35 years [1]. The methods of ML and Bayesian inference also allow to estimate TMRCA, given a specific mutation rate for the genetic region under analyses. ML also allows to specify different mutation rates for different portions of the molecule analysed in the phylogeny. Moreover, these statistical approaches allow the use of a relaxed molecular clock, which assume different rates for different branches, contrarily to a strict molecular clock tree, with the assumption of single rate for all lineages in the tree. The relaxed

molecular clock dating approach may offer better estimates of divergence times, since in many cases the assumption of rate constancy is violated. The use of relaxed molecular clocks is possible through MCMC approach that explore a weighted range of tree topologies and, simultaneously estimate the parameters of the chosen model of evolution [26, 57, 58].

1.2.3. Founder analysis method

Recently, a phylogeographic approach named founder analysis was developed by Richards et al. [59] to study non-recombining DNA markers, with the purpose of identifying and dating the migration of lineages that moved into a new territory.

The basic principle of the founder analysis is to subtract the genetic variation that has been carried by founders from the source population, during a migration/colonization event, to the diversity observed in the sink population. Therefore, only the genetic variation acquired after the colonization process is used to estimate the time since the migration event [60]. This approach formalized in the year 2000 built up on previous criticisms [61, 62] that "*The average coalescence time of two sequences sampled from two diverging populations is, in general, older, or much older, that the split of the groups. Unless a group colonizing a new territory passes through a strong and long-lasting bottleneck, part of its initial diversity will be maintained [...]"* [61].

Taking this into consideration, Richards et al. [60] developed criteria that take into account the effect of both gene flow and recurrent mutation, reducing its confounding effects, namely criteria *f*0, *f*1, *f*2 and *f*s [59]. The *f*0 criterion considers every single candidate founder as a real founder, thus it is more prone to false estimates, due to recurrent mutations and possible gene flow back to the source. However, it may prove useful to obtain maximal estimates for the most recent migrations. In order to minimize the impact of recurrent mutation and back-migrations, the *f*1 and *f*2 criteria exclude sequences that result from parallel mutations. These criteria do not allow sequence matches at the tips of the source phylogeny, which means that the founders must have at least one (*f*1) or two (*f*2) derived branches in the source population. Another criterion, *f*s, was created to account for the variation in the frequency of founder cluster candidates in the sink population, since the probability of back-migration to the source is dependent on

the frequency of the cluster in the sink population. In common clusters, the chance that back-migration or recurrent mutation will be detected is higher than in rarer clusters, therefore the criterion might be both too weak or too rigid, for common or rare clusters, respectively [59].

1.2.4. Interpolation maps

In a simple way, interpolation methods are based on the assumption of spatial autocorrelation, which states that the distance and direction between sample points can be used to estimate values at unknown points of interested. The ground behind spatial interpolation is that, on average, values at points close together in space are more likely to be similar than points further apart [63].

Several interpolation algorithms have been developed. One of the most widely used is the Kriging algorithm, which attempts to express trends suggested in the data, so that, for example, high points might be connected along a ridge rather than isolated by bull's-eye type contours [64]. Kriging is a very flexible method, thus it is very appropriated to create visually appealing maps from irregularly spaced data, such as sampled genetic data throughout a large geographical region. The genetic data can be displayed on maps, in the form of allele/haplotype frequencies within a population onto the geographical location of that population. In this sense, interpolation methods allow the integration of geographic information with genetic variation patterns, which is an interesting approach to assess the impact of the evolutionary processes that shaped the distribution of present-day genetic diversity [1].

1.2.5. Population structure methods

In the last years, a variety of methods have been developed for the analysis of population structure. These new approaches aim to infer admixture events in the history of populations, as well as to estimate a range of gene flow parameters, including ages,

proportions, and origins [65]. The most popular approaches falls into two categories: based on discrete population admixture models; or based on multidimensional statistics, such as principal components analysis (PCA) [24, 66].

1.2.5.1. Structure/Admixture

In methods based on admixture models, such as those implemented in the software packages STRUCTURE [67, 68] and ADMIXTURE [69], each individual is assumed to have inherited some proportion of its ancestry from one of several distinct populations (referred as *K*s or clusters). The key goal of admixture-based models is to estimate these ancestry proportions and the allele frequencies of each population. STRUCTURE has been used as the standard clustering method, but as it is based on MCMC methods, it is highly computably demanding when using genome-wide data [67, 68].

Recently, computationally faster methods implementing a ML based algorithm, have been developed to infer genome-wide ancestry. These include software packages such as ADMIXTURE [69] and Frappe [70]. Similar to STRUCTURE, the ADMIXTURE program analyses the probability of observed genotypes using ancestry proportions and population allele frequencies [71]. Frappe, although far more computationally efficient than STRUCTURE, is less computationally efficient than ADMIXTURE [70, 71].

1.2.5.2. Principal Component Analysis (PCA)

Another approach to infer population structure is to employ multidimensional statistic methods, such as the PCA implemented in the program SmartPCA [68]. PCA is a very useful tool for genetic data analysis, especially in the study of human migration, since it has resolution at fine geographic scales [72].

PCA can be thought as a statistical combination of a large number of measurements that accounts for the largest amount of variability in the data, which then are reduced to few principal components (PCs) that explain the main patterns [72]. It projects the individuals into a low-dimensional subspace in such a way that the locations

of individuals in the projected space reflects the genetic similarities among them [66]. Therefore, similarly to STRUCTURE, population-based PCA approach uses clusters representing discrete populations to establish genetic relationships among groups, however it is nearly as much computer demanding.

1.2.5.3. Local ancestry analysis

Global ancestry estimates, using programs such as the STRUCTURE and ADMIXTURE described above, are concerned only with estimating the ancestral proportions from each contributing population, averaged across the entire genome of an individual [73]. In the case of local ancestry estimates, the goal is to identify the ancestral origin of distinct chromosomal segments within an individual genome [69, 71]. This type of approach is especially useful in cases of admixed populations, because the genome of each individual is fragmented into shorter regions of different ancestry.

Recently, local ancestry inference has become an important method in the genetic analysis of fully sequenced human genomes. One of the current methods for inference of locus-specific ancestral information, is the RFMix [74]. RFMix is an approach that models ancestry by dividing each chromosome into contiguous disjoint windows (by using the SNPs locations) and inferring local ancestry within each window by using large reference panels. Once the ancestries have been assigned to the windows within admixed chromosomes, they are used to establish the haplotype patterns in the ancestral populations [74]. A potential limitation for this type of approach is that it requires relatively large reference panels that are good proxies for the true ancestries of the admixed populations. This limitation is being overcome with the availability of complete genomes from across the globe, as the ones being generated by the 1000 Genomes project.

1.2.5.4. Dating admixture through linkage disequilibrium decay

Most methods for genome-wide analyses (as the ones mentioned previously) are based on patterns of ancestral proportions in the genome of admixed individuals or on the divergences in the allele frequencies, but these methods do not allow to infer the time that has elapsed since the admixture events. However, since break-up of LD by recombination

is time-dependent (section 1.1.2.1.2), in theory, it is possible to infer with some level of accuracy the date of admixture from the lengths of ancestry tracts, i.e., chromosome segments with recent migrant ancestry [1]. In another words, recombination breaks down LD, leaving a signature of the time elapsed since admixture that can be inferred by LD decay [65]. ALDER is one of the methods developed to estimate the age of admixture, based on the extent of LD decay [65].

ALDER computes the weighted LD statistic for making inferences about admixture proportions and dates, with fewer constraints on reference populations than other methods (e.g. ROLLOFF [75]). One interesting feature is that it has a statistical test for admixture, i.e., it automatically determines a minimum genetic distance at which to start curve fitting (to avoid confounding signal from background LD) [65]. The LD-based statistical tests, such as ALDER, use recombination, which means that there are some limitations to detect chronologically old admixtures due to the rapid decay of the LD curve. As a result, LD-based approaches in populations with long history of continuous gene flow over time, are likely to produce date estimates that are more recent than the actual initiation of admixture [65].

1.3. GENETIC MARKERS IN HUMAN POPULATION STUDIES

The majority of the human genome is biparentally inherited and rearranged at each generation through recombination. However, two particular segments of the DNA are inherited from one parent only, and do not recombine: the mtDNA and the majority of the Y chromosome. The lack of homologous recombination results in the inheritance of these genetic systems as two haplotype blocks from one male (Y chromosome) or female (mtDNA) generation to the next, unless mutations occur [1]. These unique features make the uniparental genetic markers highly informative in terms of inferring population history, contributing for a better knowledge about the differences in female and male demographic patterns [76, 77]. Figure 5 illustrates the different patterns of inheritance between autosomes and the uniparental markers, mtDNA and Y chromosome.

Another advantage of studying haploid markers lies in the possibility of estimating the temporal scale of events, thus distinguishing different layers of migratory waves with a relatively high level of resolution. However, despite the many applications of mtDNA and

Y-chromosomal analysis, they represent only two loci in the human genome, and their effective population size is only ¼ compared to the autosomes (only one copy of these molecules is passed on to the next generation per four copies of each autosomal chromosome). All this together makes these haploid markers more susceptible to bias, whether by chance (drift) or by design (selection) [78]. Therefore inferences made solely based on haploid markers may not reflect the entire population history. In this sense, the analysis of whole genomes assumes an important role to provide a larger picture of the variation in the human genome.

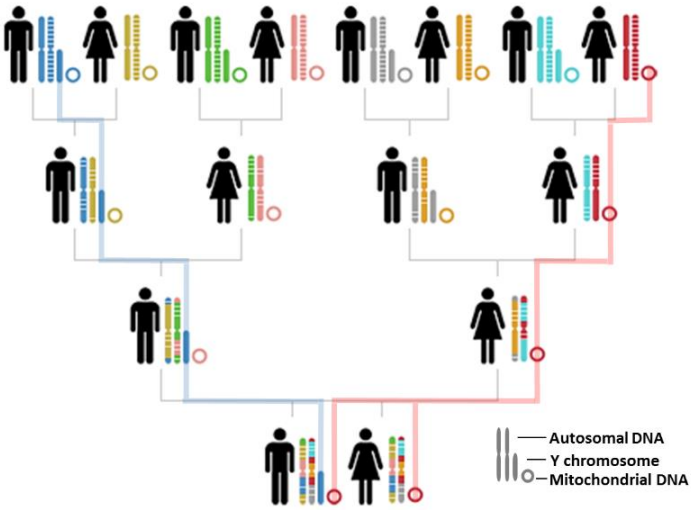


Figure 5. Transmission patterns of human genetic inheritance from recombining autosomal genetic markers and uniparental markers, Y chromosome and mitochondrial DNA. Adapted from The Genographic Project (genographic.nationalgeographic.com). The paternal lineage (NRY) passed from father to son is represented in the blue line; the maternal lineage (mtDNA) passed from mother to both daughter and son is represented in the pink line.

1.3.1. Human mitochondrial DNA

1.3.1.1. General organization

The mtDNA is organized in nucleoids within mitochondria. This organelle, present within all nucleated mammalian cells, is critical to the homeostasis of the cells. The

number of mitochondria varies considerably according to cell type and energy-request. Thus, cells in need of a lot of energy, such as muscle cells, contain thousands of mitochondria, each containing 2-10 mtDNA molecules, while other cell types may contain only a few hundred [79, 80]. The mtDNA copies can be all identical (homoplasmy) or a genetic variation can occur in a given percentage (heteroplasmy) within the cells [79].

Mitochondria are thought to descend from free-living bacteria that established an endosymbiotic relationship with proto-eukaryotic cells around 1.5 billion years ago. MtDNA preserved several of the original bacterial genome features, such as circular DNA molecules, the absence of histones, and for example, some differences in the genetic code (as previously discussed) [81]. The human mtDNA is a small circle double-stranded DNA molecule, with only 16,568 base pairs. The first complete human mitogenome was sequenced and published in 1981, and is referred as the "Cambridge Reference Sequence" or CRS [82]. Later in 1999, a revised version of the CRS, in which several sequencing errors were corrected, was published and renamed as the rCRS (revised Cambridge Reference Sequence) [83].

In terms of structure, the human mtDNA is constituted by a heavy strand (H-strand) containing a greater number of guanine bases, and a light strand (L-strand) rich in cytosine bases (fig. 6). Unlike nDNA, the mtDNA is extremely compact, with the coding region reaching approximately 93% of the total sequence. The human mtDNA contains 37 mitochondrial genes (28 on the H-strand and nine on the L-strand with no introns) that encode 13 proteins participating in the oxidative phosphorylation pathway, two ribosomal RNAs (rRNA) and 22 transfer-RNA (t-RNA) genes involved in the mitochondrial protein synthesis [5]. The mtDNA sequence of some genes (e.g. ATP6 and ATP8) overlap and, in other cases, the genes are contiguous or separated by just one or two non-coding nucleotides [82]. Apart from the extensive coding region, a segment with regulatory function (~1.2 kb), forms a non-coding region, referred as control region or D-loop (i.e. displacement-loop), extending from positions 16,024 to 576 [5]. The control region contains two hypervariable regions (HVRI and HVRII) ranging from positions 16,024-16,383 and 73-340, respectively, with a higher mutation rate than the rest of the molecule. This mtDNA region has been extensively used in population genetic studies, due to its high discriminating power.

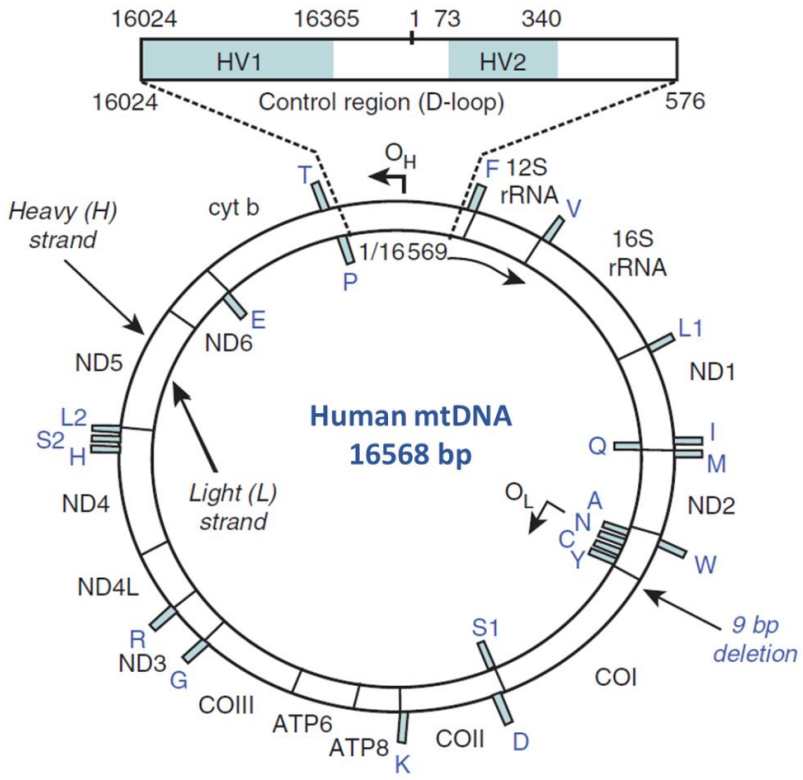


Figure 6. Schematic human mtDNA. The H-strand is represented by the outside line and the L-strand the inside one. The abbreviations in the figure correspond to: ND1-6: NADH dehydrogenase subunits; COXI-III: Cytochrome c oxidase subunits; ATP6 and ATP8: subunits of ATP synthase; and cyt b: Cytochrome b. The non-coding control region, representing a stranded D-loop, is shown in more detail at the top of the figure. Adapted from [5].

1.3.1.2. mtDNA unique features: mutation rate and mode of inheritance

The mtDNA mutation rate is not uniform throughout the molecule, with comparative low rate in the coding region, and a considerably higher rate in the control region (especially in HVRI and HVRII). Mutation rates have been the focus of several studies that employed different methods, including family pedigree [84, 85] and human phylogeny approaches (either limited to HVR region [55], or including the overall coding region [86] or just based on synonymous mutations in the coding region [87]). Substitutions rates have also been estimated with distinct calibrations points, such as the divergence between chimpanzees and humans [86, 87], and estimated date of human phylogeny (branching points in the tree) calibrated with archaeological, biogeographic and/or climatic information [88-90].

In the last years, studies have shown the importance of selection (more precisely purifying selection) on the mtDNA genome [91]. Purifying effect leads to the decay of nonsynonymous mutations over time, meaning that younger branches in the mitochondrial phylogeny will have a higher proportion of non-synonymous mutations than older branches, which may result in an overestimation of the age of younger lineages. Therefore, the use of an average molecular clock over all mtDNA positions may be inadequate.

Taking in consideration the fact that the mutation rate is time-dependent in the mtDNA, Soares et al. [91] estimated the effect of purifying selection on the coding region of more than 2.000 mtDNA complete genome sequences, and proposed an improved molecular clock for dating the whole human mtDNA genome, incorporating both coding and non-coding regions. The new improved mtDNA molecular clock was calibrated without any prior assumptions on intraspecific calibration points and against recent evidence for the time of the *Homo-Pan* split, and as expected, the estimated coalescent times showed that the more recent branche ages obtained with the previous linear mutation rates were slightly overestimated. For the complete mtDNA genome, an estimate of 1.665×10^{-8} ($\pm1.479\times10^{-9}$) substitutions per nucleotide per year (i.e., one mutation every 3624 years) was obtained. The rate of synonymous mutations was also estimated, and corresponds to one substitution every 7884 years.

In addition to the high mutation rate (overall ten times higher that of the nDNA), mtDNA has a unique mode of inheritance, another valuable feature for evolutionary genetic studies. As mentioned before, it is maternally inherited as a single unit from generation to generation. This lack of paternal transmission could result from either (i) the dilution effect, because the unfertilized egg contains ~100,000 mtDNA molecules, (ii) the removal of the sperm mtDNA by ubiquitination or (iii) the occurrence of a “mtDNA bottleneck” excluding the paternal alleles [92]. There was only one example identified of paternal transmission, possible due to failure of the mechanism for recognition of paternal mitochondria [93]. Since, in general, the mtDNA is inherited uniparentally, through the maternal germ line, both females and males inherit their unchanged mtDNA (excluding for mutational events) from their mother but the males cannot transmit it to subsequent generations.

1.3.1.3. mtDNA phylogeography

Due to the high mutation rate of mtDNA, the variants accumulate fast enough amongst different geographic locations, making this genetic system a powerful and widely used tool for studying demographic events and migratory occurrences [79]. The mtDNA haplotypes are classified in different groups, sharing the same point mutations, designated by haplogroups [94]. The nomenclature of the haplogroups follows the branching structure of clades and subclades within the tree: haplogroups are initially designated with capital letters and the derived subclades are named intercalating lower-case letters with numbers. Additionally, the star symbol (*) is used to define members of a haplogroup that do not belong to a defined sub-clade [94].

Early studies have appointed an African origin of the mtDNA molecule, and the human population itself. In 1987, Cann et al. [95] with a worldwide population study introduced the “Mitochondrial Eve” concept, which argues that all present-day mtDNA traces back to an unique ancestor who lived in Africa, probably around 200 thousand years ago (ka). In light of this concept, since mtDNA is maternally inherited, the variation present in the mtDNA derives from an unique African female ancestral lineage, which should not be confused as an unique living individual, but an ancestral population. mtDNA phylogenetic patterns showed that when populations started to growth and expand, mutations started to accumulate in the mtDNA lineages that colonized different geographic regions, to a point where haplogroups became continent-specific.

As shown in figure 7, the oldest branches in the mtDNA tree are restricted to African lineages, in fact, the first six splits in the mtDNA tree define diverging strictly sub-Saharan African branches (L0-L6) and the one that harbours Africans and all non-Africans branches (L3). The rare African lineages found in Europe, West Asia and America have been mostly associated with human trade during the Roman times, Arab conquests and Atlantic slavery [96, 97]. Haplogroup L3 arose, mostly likely, in east Africa where it dispersed *in situ*. In the late Pleistocene (~60 ka) carriers of L3 made their way out of Africa, giving rise to all non-African mtDNA diversity which is affiliated in M and N super-haplogroups. The number of non-African macrohaplogroups can, however, be extended to include a third one within the daughter-clade haplogroup N, named haplogroup R. Macrohaplogroups M and N (including R) gave rise to several subclades with a specific geographic distribution. Haplogroups U, HV, JT, N1, N2 and X are frequently found in North Africa, Europe and Southwest Asia; and haplogroups A, B, F, O, P, S, T, M and Y

are widespread in Asia, Oceania and Native American populations [49]. Interestingly, Asian haplogroups A, B, C and D are considered the pioneers who populated the Americas through a migration across the Beringia Strait [94].

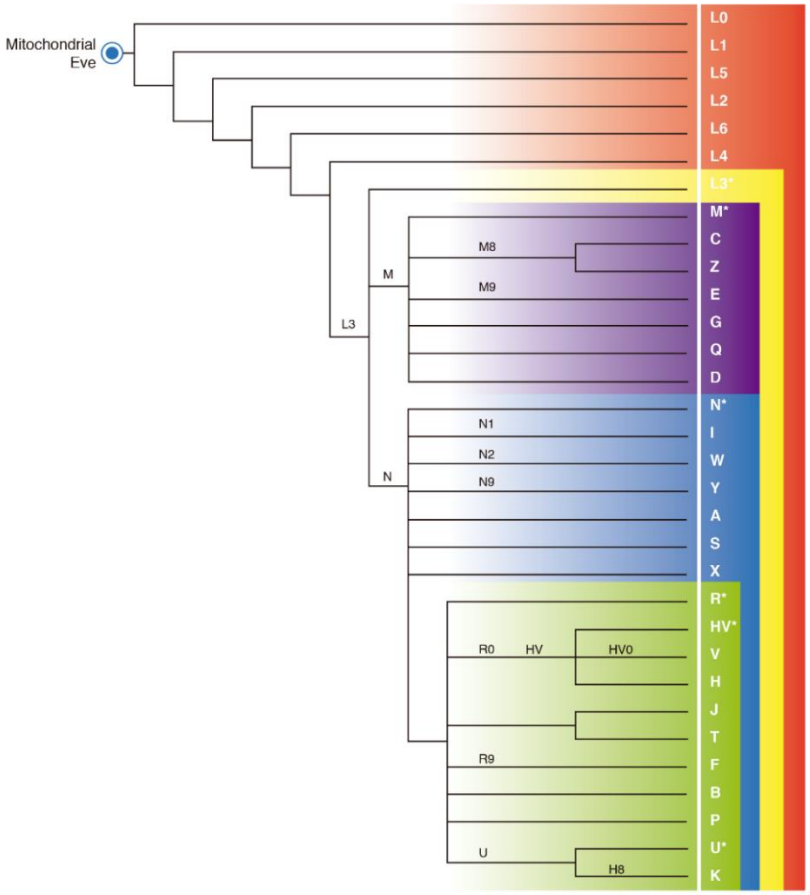


Figure 7. Schematic global simplified human mtDNA phylogenetic tree. Adapted from [98].

1.3.1.4. Dispersal of the first modern humans

Recently, with the development of high throughput technologies, the sequencing of complete mitogenomes has become common, allowing to obtain a better resolution than the one obtained only by sequencing the complete control region. Complete mtDNA studies confirmed that African populations present a higher genetic diversity than non-African populations, as a result of their longer genetic history. On the other hand, non-African populations, present a “star-like” phylogeny, suggesting a recent bottleneck event, and subsequent strong population expansion [99]. These conclusions have been

considered the foundation stones of the out-of-Africa hypothesis. According to this hypothesis, the first modern humans arose in Africa around 200 ka. After ~70 ka, with the climate improvement, the humans populations started to grow, and expanded from sub-Saharan Africa to the rest of the world, interbreeding to a certain extent and largely replacing other existing populations/species as Neanderthals in Eurasia. Several out-of-Africa routes have been proposed, but two have received more support: a northern migration route through the Levant, towards Europe and Asia, around 45,000 years ago; and a southern route, from East Africa, through the south coast of the Arabian Peninsula towards Asia, around 60,000 years ago [100]. Theoretically, the first modern humans could have spread in several directions simultaneously, however, the southern coastal route towards the east appears to be sufficient to explain the existing mtDNA variability, not only in North and Southeast Asia, but also in Oceania [101].

The phylogeography of mtDNA variation - and also Y chromosome variation (that will be discussed further) - supports this single successful exit from Africa, most probably after the Toba eruption (~74 ka) [101]. There is archaeological evidence that older out-of-Africa migrations occurred, namely through the Levant, but these were not successful in leaving descendants till the present [102]. There are three main mtDNA evidences supporting the out-of-Africa unique successful migration. First, all non-African mtDNA lineages descend from two haplogroups M and N, which diverged from sub-Saharan L3 lineages, most probably in India and the Gulf region, respectively [101]. Second, the new improved molecular-dating techniques estimated the origin of L3 within Africa around ~60-70 ka (thousand years ago) [90] and a close coalescent time for its non-African daughter clades M and N (60 to 50 ka) [91], which suggests that most probably all the mtDNA gene pool outside of Africa had an origin on a single rapid out-of-Africa migration event [90, 103]. And third, different geographic regions of South, East and Southeast Asia, and Australia harbour different lineages descending directly from the root of the basal non-African branches M, N and R. The same goes for the Near East and Europe, who also present specific N and R mtDNA lineages, but do not harbour the M lineages [104-113]. This further support that out-of-Africa founders dispersed along the south coast of Asia, reaching Southeast Asia and the Pacific around 60 ka, and only later reached Europe, carrying only two of the founders, N and R [97, 114].

After the out-of-Africa, the Eurasian founders have undergone substantial demographic change and additional migrations, particularly, in the Late Glacial and postglacial periods. For instances, the majority of the present-day European mtDNA haplogroups appear to descend from Palaeolithic pre-agricultural populations and only

minor lineages were brought by later-expanding farmers. This suggests that after the African exodus, during the cold periods (Last Glacial Maximum – LGM, and Younger Dryas) the Palaeolithic populations took refuge in southern refugia, and later, with the climate improvement, re-populated Europe. The re-expansion of populations from the Franco-Iberian refugial areas led to the dispersal of common European mtDNA lineages H1, H3, H5 , V and U5b, whereas the descendants of the inhabitants from the Italian Peninsula and the East European Plain lead to the spread of the maternal lineages U5b3, and U4 and U5a, respectively. European haplogroups I (within N1a1b), J, T and W also appear to have been carried by Palaeolithic populations that expanded from Near Eastern glacial refuges [115-121]. The later spread of agriculture and pastoralism from southeastern Europe into central Europe and the Mediterranean, appear to have resulted into indigenous dispersals carrying some of these Palaeolithic lineages (e.g., several haplogroup H subclades) [122]. Another important demographic event was the peopling of the Americas probably at the end of Pleistocene, in which populations from Asia (Siberia) dispersed through the Beringia Strait into the New World, carrying with them the four major pan-Americans haplogroups A, B, C and D, and one minor North American haplogroup X [94, 123].

The mtDNA analysis is a powerful tool to estimate biogeographical ancestry at a broad geographic level. But this genetic system only accounts for the maternal variation in the populations, thus it might not reflect the complete population history. Therefore, it is important to complement the studies based on mtDNA variability with other markers, such as Y chromosome and autosomal data [79].

1.3.2. The Y chromosome

1.3.2.1. General organization

The Y chromosome is the third smallest chromosome within the human genome, with approximately 60 Mb. The complete sequencing of the human Y chromosome was published in 2003 [124]. The Y chromosome contains five distinct types of sequences (fig. 8): heterochromatic, X-transposed, X-degenerate, ampliconic and pseudoautosomal

regions (PAR), these last recombine with X chromosome homologous regions, and are located at the extreme termini of the Y and X chromosomes [125, 126].

The non-recombining region of the Y chromosome (NRY), recently renamed as MSY (male specific region), accounts for 95% of the Y chromosome's total length [5]. The MSY is mainly responsible for the male sex determination.

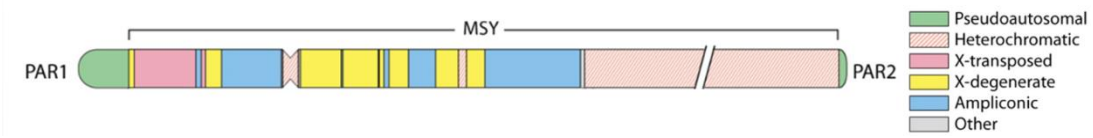


Figure 8. Schematic Y chromosome with the heterochromatin and pseudoautosomal regions identified. Adapted from [126].

1.3.2.2. Mode of inheritance and mutation rate

Similarly to mtDNA, the Y chromosome is an haploid marker inherited through the paternal line. The NRY, usually, passes intact from generation to generation, descending as a single locus, in a way that the haplotipic diversification occurs through the accumulation of mutations in time, preserving a simple record of their story. Furthermore, the effective population size of the NRY is equivalent to that of the mtDNA, which means that in a population it represents one quarter of that of any autosome, and one third of the X chromosome. Because of this reduced effective population size, the Y chromosome genetic variation is more susceptible to genetic drift that modifies the frequencies of different haplotypes, especially in small populations. The increased genetic drift also leads to large population clustering and further differentiation of NRY SNP loci, rendering the Y chromosome one of the most geographically informative loci in the genome [1].

The Y chromosome has become a useful haplotype system in studies of human migration and evolutionary history, through the identification and characterization of microsatellites or STRs, and bi-allelic markers (SNPs) [126]. Microsatellites or STRs in the human Y chromosome are highly polymorphic between individuals, making them very useful in the forensic field for individual identification. In the case of Y-SNPs, they have a similar mutation rate to the autosomal SNPs, approximately 10^{-8} per base pair per

generation, which is considerably slow compared to the Y-STRs [127]. Because of their low mutation rate, each SNP is consider an unique and independent event in evolution. Therefore, the study of SNPs in the Y chromosome allows to easily reconstruct their phylogenies, providing invaluable information regarding male large-scale population dynamics [128, 129]. The higher mutation rate of Y-STRs, makes them very useful to determine intra-haplogroup diversity and to attribute age estimates to MRCA, however there is considerable uncertainty in STR mutation rates estimated through pedigree and phylogenetic methods. The pedigree mutation rate yields an average of $2.0x10^{-3}$ per generation, while the evolutionary mutation rate is estimated to be approximately $2.6x10^{-4}$ per generation for the same STR loci [127, 130]. One possible reason for this difference is that pedigree methods detect new mutations which are lost through genetic drift in populations, thus not been detected in the phylogenetic methods [131].

1.3.2.3. Y chromosome phylogeography

Two decades ago, several researchers started to study, in parallel, the binary Y chromosome polymorphisms, in many different populations to characterise paternal genetic patterns and to make inferences on human evolution, ethnic affiliations and general demographic history. In order to deal with the constant increase in the number of Y-SNPs, the Y Chromosome Consortium (YCC) developed a universal nomenclature system for the human Y chromosome tree, publishing the first highly resolved phylogenetic tree of binary Y chromosome haplogroups in 2002 [132]. The phylogenetic structure was determined by comparing homologous MSY of closely related species (e.g., chimpanzees, gorillas and orangutans), which provided an outgroup for the tree rooting, thus allowing to determine the likely ancestral state for human polymorphic sites [133]. At that point, the phylogenetic tree, was based on 245 Y-SNPs, defining a total of 153 binary-defined haplogroups. The major clades of the phylogenetic tree were represented by 18 haplogroups, starting with the haplogroup A, and then in alphabetical order until letter R [134].

The Y chromosome parsimony tree was later updated by Karafet et al. [128] (fig. 9), comprising 599 binary markers. The Y tree has been maintained ever since, in the interactive Webpage “snpreferencedatabase” (http://www.snp-y.org/), to avoid

nomenclature problems, and to validate the application of the Y-SNPs in population studies.

The overall Y chromosome haplogroup distribution is phylogenetically well structured. Haplogroups A and B are the oldest branches, the first mostly present in hunter-gatherer populations of Ethiopia and Sudan, while haplogroup B is more frequently found in Pygmies [78]. The major DE and CF branches contain the remaining non-African Y-chromosome haplogroups (C to T). Haplogroup D presents the highest frequencies in Tibet and Japan, and is also found throughout Central and Southeast Asian regions, while its sister clade E is highly frequent in African populations, as well in some Mediterranean and European populations. Haplogroup C is found in Asian, Oceanic and Australian populations, and one of its subclades, C3b, is restricted to Native American populations [128, 133]. Haplogroup F is the ancestral of the remaining G to R clades: paragroup F* and haplogroups F1 to F4 and H are restricted to Asia populations; G appears with more frequency in the Caucasian region, the Near East and in the Mediterranean areas; I lineages are restricted to Europeans populations, contrary to its sister clade J, which is frequent in the Near East, North Africa, Europe, Central Asia, India and Pakistan [135]. Haplogroup K is the last major branch of the Y chromosome tree, encompassing haplogroups K to T: K lineages are widely distributed all over the world but at low frequencies; haplogroup L is mainly found in the Indian Subcontinent, and, occasionally, in European populations, especially in Mediterranean countries; haplogroup M presents highest frequencies in Oceania and Eastern Indonesia; haplogroup N is found in Northern Eurasia, probably originated in the south of Siberia; O is the most predominant haplogroup within the region of East and Southeast Asia; clade PQR encompasses lineages belonging to minor frequency haplogroup P found in Caucasus and India, haplogroup Q, found in Europe, Northeast Asia and in the American continent, and haplogroup R mainly found in Europe [133]. The last two haplogroups newly defined by Karafet [128] are haplogroup S and T (previously denominated K5 and K2, respectively), the first is frequent in Oceania and Indonesia; and the latter is found at low frequencies in the Near East, Africa and Europe [128].

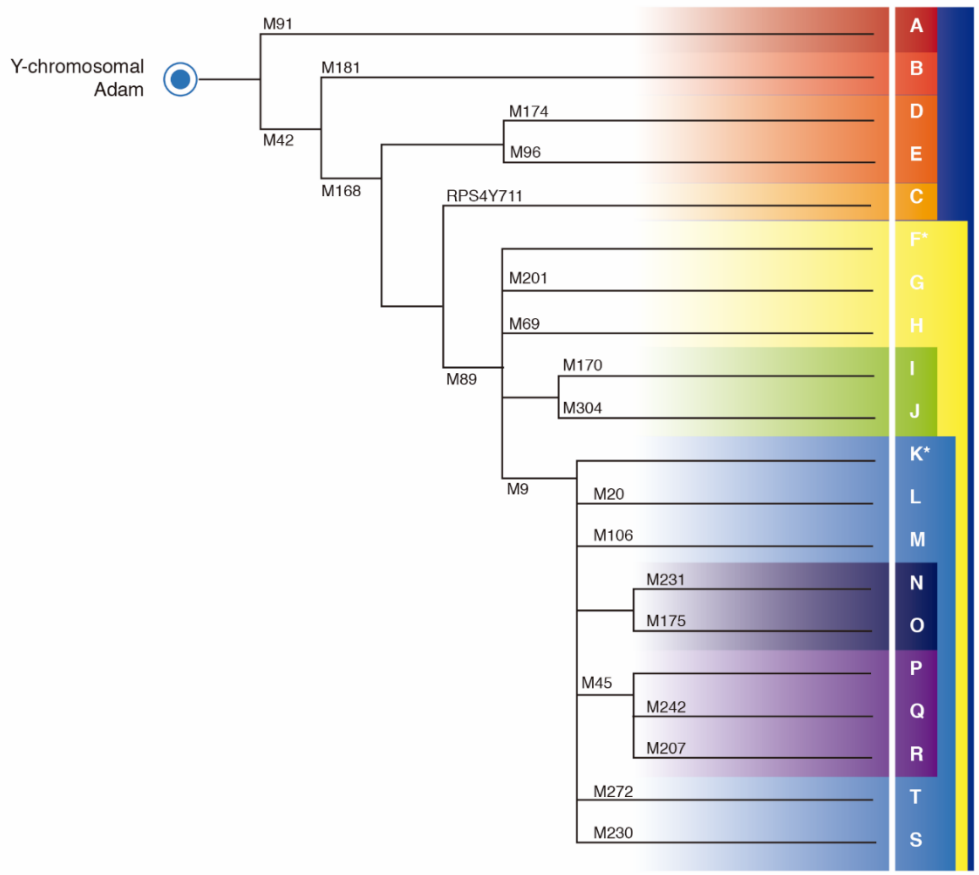


Figure 9. Schematic global simplified human Y chromosome phylogenetic tree. Adapted from [128].

1.3.2.4. Y chromosome evidences of the out-of-Africa

Similarly to the mtDNA, the global phylogeographic distribution of Y chromosome haplogroups is consistent with an African origin of the modern human species, thus supporting the out-of-Africa hypothesis, with all non-African haplogroups deriving from the sister-clade B [126].

Though there is substantial uncertainty in dating Y chromosome, it is estimated that MRCA of all extant human MSYs lived during the interglacial period approximately between 130 to 90 ka (nicknamed as the Y-chromosomal Adam). Around 50 to 40 ka, the out-of-Africa event took place.

After the initial settlement of Eurasia, it appears that small groups of modern humans, holders of the founders C, F and K clades, started to split into several isolated

groups, developing region-specific genetic variants [78]. At the end of the LGM, with the improvement of the climate conditions, populations started to expand from glacial refuges in Franco Cantabria and East Europe and/or Balkans regions, carrying with them several clades, as R1a and R1a1 [136]. The beginning of the Neolithic in the Near East, was another main impulsion for the demographic expansion, leading to de dispersal of several lineages within I, G and J2 haplogroups [78, 137].

1.3.3. The autosomal nuclear genome

1.3.3.1. General organization

The human karyotype contains 22 pairs of autosomes (non–sexual chromosomes), that can be identified by their size, banding pattern and centromere position. They are organized from the largest (chr 1) to the smallest (chr 22), with the exception of chromosome 21, which is, actually, shorter than chromosome 22 (fig. 10) [1].

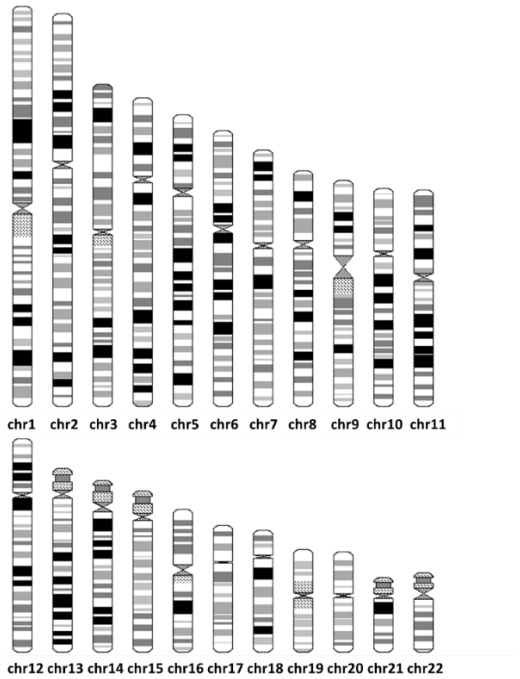


Figure 10. The human karyogram. Adapted from The International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/karyogram/gwas.html). The chromosomes are arranged by size. Normal human cells (except gametes) contains two copies of each autosome.

1.3.3.2. Mutation rate

The average mutation rate for autosomal SNPs was estimated to be approximately $2x10^{-8}$ per generation, which means that the probability of the same change occurring twice independently is very unlikely [138]. This is an important feature in genetic studies, because it allows to infer human ancient history, but it is not so informative in terms of recent history. Since random evolutionary forces affecting two distinct populations cause differences in the variant frequencies, these differences might be very useful to infer, for example, the probable origin population of a group of individuals [1].

1.3.3.3. Development of genome-wide analysis

A landmark event in nDNA analysis was the release of the first nearly complete sequence of the human genome in 2004 [139], under the initiative of the Human Genome Project (HGP) initiated in the 1980s [140, 141]. By early 2000, other large-scale genetic projects were established, such as the Human Genome Diversity Project (HGDP) [142] and the HapMap [143]. The HGDP, with a more anthropological view, intended to describe the global human genetic variation, therefore providing a database for research on human genetic migratory history [144]. On the other hand, the HapMap project, started as an international group involving United Kingdom, United States, Canada, China, Japan and Nigeria, with a biomedical approach and aiming to describe common genetic variation in human populations [143]. The HapMap took advantage of the non-random association of common variants found on large chromosomal sections (LD) to map disease genes [145, 146]. Taking into consideration the LD, it became possible to catalogue human genetic variation by genotyping specific SNPs in each haplotype region of the genome.

Other recent genotyping projects for the characterization of human genomic variation, include the Encyclopedia of DNA Elements (ENCODE) [7, 147] and the 1000 Genomes Project [148]. The ENCODE Project was established with the goal of identifying all functional elements in the human genome sequence. In order to achieve that, regions of transcription, transcription factor association, chromatin structure and histone modification, have been systematically mapped at high resolution, providing new insights into the organization and mechanisms of gene regulation [147]. The 1000 Genomes Project was the first project to sequence the genomes of a large number of people, with

the goal of developing a comprehensive resource of human genetic variation across worldwide populations. This project set out to find most genetic variants that have frequencies of at least 1% in the populations studied, which so far belong to five super populations: African; European; Admixed American; East and South Asian [148, 149].

This big amount of information available on worldwide human diversity led to the design of SNP chips (mainly commercialized by Affymetrix and Illumina), that contain thousand and millions of SNPs distributed along the genome, allowing its characterization in a single assay [150]. Together all these large-scale projects and technologies fomented the development of advanced molecular biology techniques, such as linkage disequilibrium studies, and more recently, genome-wide association studies (GWAS), which enable researchers to use common genomic variants to map multiple genes that influence common complex traits [151].

1.3.3.4. Genome-wide data in population studies

With the advance of new sequencing methods autosomal data has become a powerful tool in population genetics studies. Early autosomal DNA studies were concentrated on the analysis of STRs or microsatellites [152]. Although STR studies provided some important insights into human population history [153-155], they have been largely replaced by genome-wide SNP studies.

Genome-wide data provides information of many independent loci, and so, it offers more reliable insights into population history. Genome-wide SNPs are particularly useful for inferring past population demographic history, such as population size changes, migration events, as well as the effect of selection on human adaptation to new environments and the overall genetic structure of admixed populations [24, 156]. In particular, the level of admixture within the population is important not only in evolutionary studies [153, 157, 158], but also in the discovery of disease-associated genes [159, 160].

In the last years, several population studies started to emerge taking advantage of the bioinformatics methods that have been developed to harness genome-wide data and to infer demographic population events (discussed in 1.2.5). Studies based on genome-wide variation in populations from all over the world showed that individuals from the same or close populations evidenced similar ancestry proportions. In particular, Li et al. [161] demonstrated that it is possible to detect correlation between genome-wide variation

and local environmental and/or phenotypic variation, so that genetic structure is detected not only between different continents, but also among sub populations within continents (fig. 11A). Moreover, the results supported the out-of-Africa model of modern human origin, as the African populations contain the highest genetic levels of diversity (fig. 11B) [161].

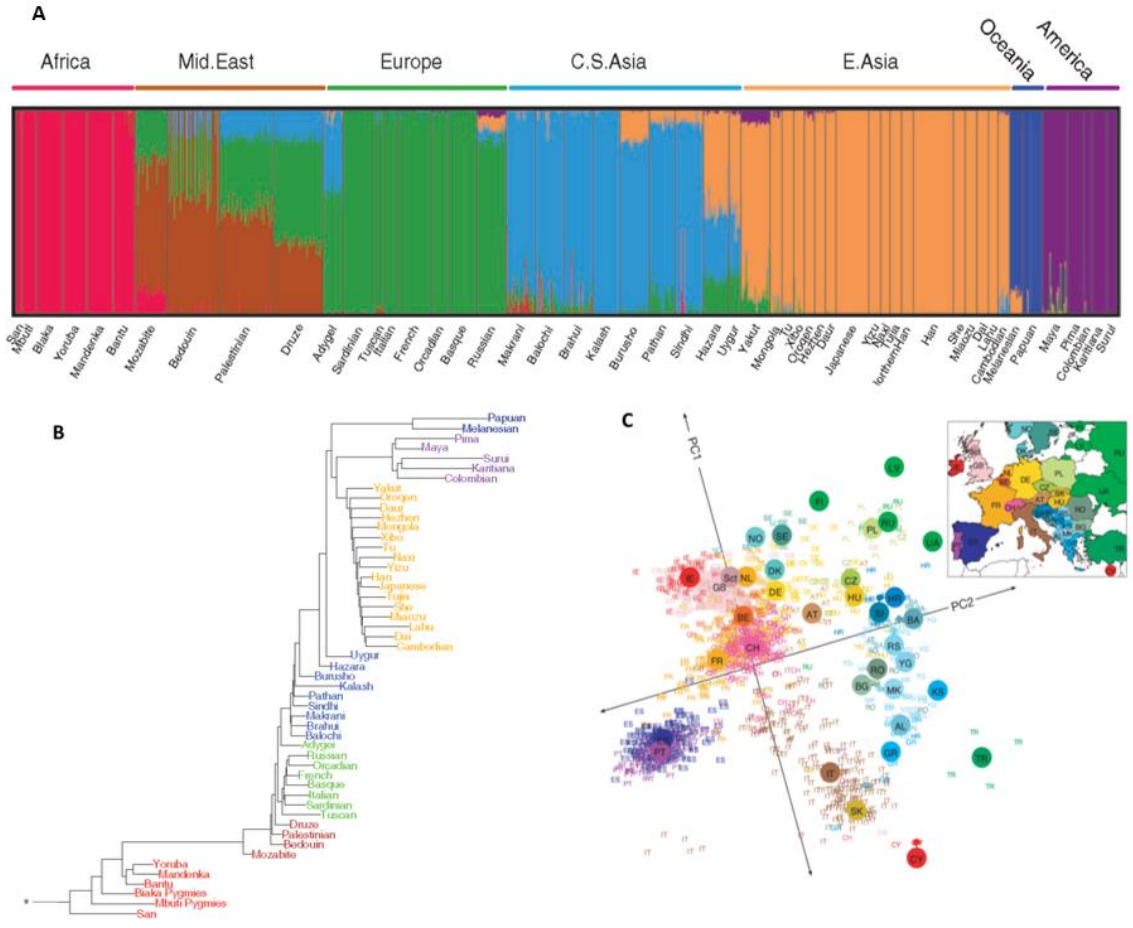


Figure 11. Patterns in human population structure. (A) Global ancestry inferred with the maximum likelihood based frappe software, using seven inferred ancestral groups. (B) Maximum likelihood tree, with the sub-Saharan African populations located nearest to the root. (C) PCA statistics from 1,387 Europeans, showing a close similarity between the genetic and geographic map of Europe. Illustration A and B are from [161] and C is from [162].

The study performed by Novembre et al. [162], revealed the potential of using genome-wide data in combination with extensive geographic sampling, for characterizing the genetic structure at high resolution on a local scale. These authors used PCA to

reconstruct the genetic variation in Europeans. The results showed that, though Europeans show little overall genetic differentiation between populations, they display a surprisingly correspondence between genetic variation and geography within the continent. In fact, the results of this study provided a two-dimensional genetic map of Europe that almost overlapped with the geographic map itself (fig. 11C).

These results are part of a long line of studies using large-scale human genome databases, aiming to unravel new insights into population structure. The few genome-wide SNP studies performed in Southeast Asian populations, to date, will be discussed later on this chapter.

2. SOUTHEAST ASIA POPULATION PREHISTORY

2.1. GENERAL FEATURES OF SOUTHEAST ASIA

Southeast Asia represents the subregion of Asia that is geographically situated south of China, east of the Indian subcontinent and northwest of Australia. It comprises the countries extending from Myanmar to Indonesia (with the exception of West Papua). The most significant geographical division has been between two main regions, roughly defined as Mainland Southeast Asia (MSEA) and Island Southeast Asia (ISEA). MSEA, also known as Continental Southeast Asia or Indochina, includes the countries of Myanmar, Thailand, Laos, Cambodia, Vietnam and Peninsular Malaysia. The region of ISEA, also referred as Maritime or Insular Southeast Asia, comprises the countries in the maritime region of SEA, namely, East Malaysia, almost all of Indonesia (excluding West Papua), Brunei, Singapore, East Timor, Christmas Island and Philippines. Although Taiwan is sometimes consider as part of ISEA [163], throughout this thesis I will refer to Taiwan separately.

In terms of biogeography, SEA can be divided by the so called Wallace's line, into two regions with distinct biotic contexts, namely Sunda and Wallacea (fig. 12) [164]. The Wallace's line can be seen as a deep-water channel that marks the southeastern edge of the Sunda shelf which links Borneo, Bali, Java, and Sumatra underwater to MSEA. In the

same way, Australia is connected by the Sahul shelf to New Guinea [163, 165, 166]. The Sundaland region, since it corresponded to a mainland extension of dry land existent in the Late-Pleistocene, has typically continental Asian biota, whereas the region of Wallacea, which has never been linked to the mainland, and was only populated by organisms capable of crossing the straits between islands, presents a considerably impoverished island biota [167].

Another possible division used in biogeography is within the Sunda shelf itself, and represents the separation between the Indochinese and Sundaic provinces, the first includes MSEA (excluding the Malay Peninsula), South China and Taiwan, and the second corresponds to the ISEA part of Sundaland, the Malay Peninsula and Palawan in the Philippines [168, 169].

Figure 12. Map of Southeast Asia and Taiwan. It shows both the modern coastlines (dark shading) and the 120-m depth contour below sea level (light shading), indicating the extent of Sundaland, Wallacea and Sahul at the LGM.

2.1.1. Climate changes at the end of Ice Age and the drowning of Sunda shelf

The region of Southeast Asia contains a wide topographical diversity, with coastal plains, highlands and river valleys throughout all the region. This region harbours a tremendously diverse natural vegetation, such as tropical rainforests, savannah woodlands and grasslands, montane forests, and many other types of vegetation [170, 171]. SEA vast biodiversity is attributed, in part, to the region's geographic position in the humid tropics, its history of dramatic landmass changes, and its habitat division [170].

In the Late Pleistocene, SEA was subjected to multiple oscillations of warmer and colder periods. During the longer cooler glacial phase around 15,000 to 25,000 years ago (LGM), the sea levels were considerably lower, about 120 meters below the present value [172]. With the climate warming at the end of Ice Age, the immense glaciers that covered the northern hemisphere melted away, leading to three major episodes of sea level rise. The first from 13,500 to 15,000 years ago, followed by the a second flood around 10,000 to 11,500 years ago, and, probably, a third one in the Holocene about 7,000 to 8,000 years ago [165, 173-175]. The rise of the sea levels end of the LGM had a profound impact in the region, with the loss of almost half of the land area of the continent of Sundaland, and a concomitant doubling of the length of the coastline, which ultimately resulted in the current maritime division between MSEA and ISEA. These changes, most probably, trigged drastic population displacements in the region, and it probably had an important impact, in particular, on the development of the maritime culture and sailing technology seen in Sunda populations [165, 172, 176].

2.1.2. People and languages

Traditionally, SEA populations used to be divided into the so called Negritos and the Mongoloids (noteworthy, the term mongoloid is now considered pejorative by most anthropologists). The Negritos, with a characteristic dark skin and small stature, are found within SEA in the Philippines and Malaysia (Semangs). The Negritos are hunter-

gatherers, and their Australoid physical features are considered residual elements of the first Australo-Malenesian settlement [163].

According to the most common view, the light-skinned Mongoloid populations encompass almost all the populations in Southeast Asia. Southern Mongoloid Austronesian speaking peoples are the ancestor of most present-day ISEA populations, whereas mainland populations are descendants of Southern Mongoloid Austroasiatic speakers, who gradually expanded southward [163, 177]. In terms of physical features, the Mongoloids are generally short or medium stature with a yellowish skin. These two population groups often intermixed to give rive to phenotypically intermediate populations, such as the Senoi from Malaysia [163]. The majority of Mongoloid populations practice an agriculturalist mode of life, however, some Mongoloid populations in Sumatra and Borneo practice a hunter–gatherer lifestyle. According to Bellwood, these Mongoloid populations are probably farmers who went back to the forests [163].

There are four main language families spoken in Southeast Asia: Austronesian, Austroasiatic, Tai-Kadai and Sino-Tibetan. The distribution of these language families is shown in figure 13. The Austronesian language is by far the most common language in ISEA. It is also spoken in the north coast of New Guinea, the islands in the Pacific and from Madagascar to the east of the African continent. In fact, it is the language family with the widest distribution in the world, estimated to comprise nowadays up to 1,200 identifiable languages [177-179].

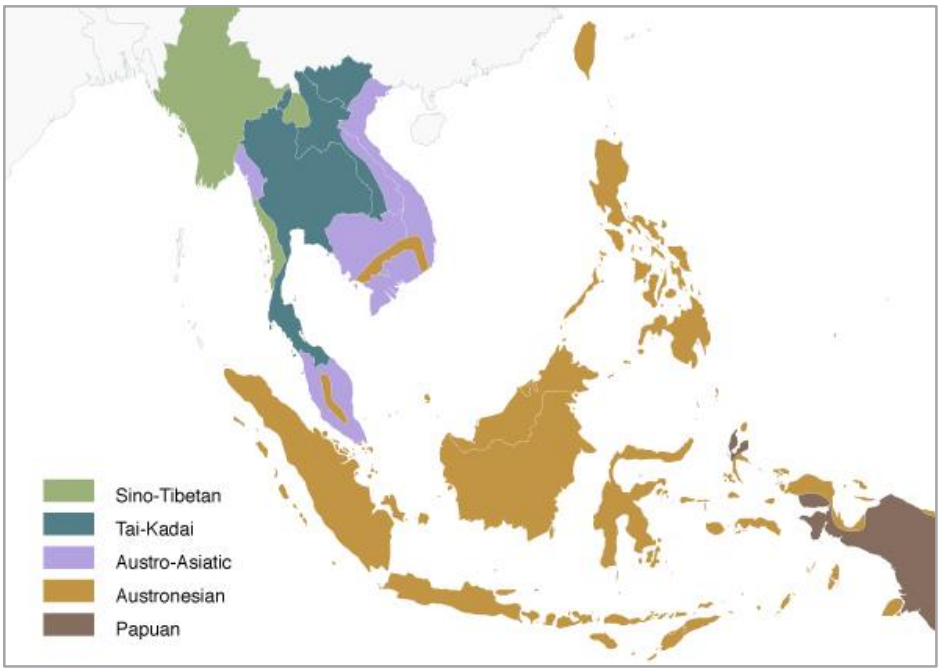


Figure 13. Ethnolinguistic distribution in Southeast Asia. Source [177].

MSEA shows a more complex linguistic pattern. The major group of languages belongs to the Austroasiatic language family, comprising about 150 separate languages, spoken by Vietnamese, Cambodians, and some hill peoples of Northern Myanmar, Assam and Laos. Its speakers also include most of Malay and Nicobarese aboriginals. The Austroasiatic family also comprises the Munda languages from eastern regions of India, such as Bihar [177]. The third most common language family is the Tai-Kadai, which is widely spoken in Thailand, and by some populations in Myanmar and the lowland region of Laos, north Cambodia, Vietnam and Malaysia. Languages from the Tai-Kadai language family are also spoken in certain borderland regions of China. The last common language family is the Sino-Tibetan, to which the Chinese language belongs to. One of its major language branches is the Tibetan-Burman, spoken by the lowland Burmese and certain hill groups in Myanmar and neighbouring mainland countries, such as Southern China, India and Bangladesh [163, 177].

2.1.3. Cultural transformations in South China and Taiwan

Early archaeological evidences of rice farming discovered in Neolithic sites in Guangxi, Yunnan, and Chongqing shed some light on the origin, development, and dispersal of early rice agriculture in southern China [180, 181]. Subsistence evidences suggest that, between 10,000 to 16,000 years ago, hunting and gathering was the major food-seeking strategy in Southern China [182]. Hunted animals included pigs, deer, birds, and riverine food resources, such as fish, freshwater turtles, and shellfish. There were also evidences of some seeds and few rice remains [181].

This cultural picture started to gradually change in the Neolithic Southern China (7,000-9,000 years ago) due to the development of a sort of food producing subsistence system. This subsistence system was marked by the increasing rice cultivation in an undomesticated form and the initial domestication of pigs and chickens; however, evidences suggest that various aquatic and non-cultivated forest plants (e.g., water caltrop, lotus and oak), along with wild animals, were still important food resources of the prehistoric societies in this phase [181, 183]. The increasing number of sites dating to this time frame, in the middle and lower Yangtze basin, indicate rapid population growth in the region. It has been suggested that the population increase along with the local absence of

raw materials for manufacturing stone tools led to the development of structured exchange systems [180, 181, 183]. The same cultural transformations emerged in Taiwan, where the number of local middle Neolithic sites multiplied considerably. Agriculture became well developed and lithic and jade stone artefacts were exchanged extensively in Taiwan Strait [181, 182, 184].

Around 4500-5500 years ago social structures and settlement patterns in the middle and lower Yangtze basin changed dramatically, becoming more hierarchical and complex. This shifting to intensive social and cultural development and population growth resulted in the full blossom of agriculture in Fujian/Taiwan and southwest China, that subsequently led to the second phase of population outflow, this time into MSEA and ISEA [181, 182]. The cultural expansion from Taiwan into northern Philippines and ISEA marked the spread of Austronesian languages, and in opposing direction, the spread of farming into Northern Vietnam and Thailand was accompanied by the dispersal of the Austroasiatic languages [185, 186].

2.2. POPULATION MOVEMENTS IN SOUTHEAST ASIA

The complex history of SEA populations is often framed by an initial settlement of Australo-Melanesian peoples out of Africa, followed by a second wave of inhabitants from East Asia, including Southern China and Taiwan, during the agricultural expansion of the Neolithic. Although, most studies of the populations movements in the region tend to focus primarily these two widely accepted migrations, in the present thesis the main proposed population movements in the region of SEA are further described in a chronological order.

2.2.1. Initial settlement of SEA by modern humans

The arrival time and migration route of the first modern humans in SEA from Africa is still controversial [91, 187-189]. There are two main opposing models for the initial peopling of SEA. In the first model, modern humans dispersed from Eastern Africa across

the Red Sea reaching SEA around 50,000 to 60,000 years ago [90, 190-192]. The proponents of the second model argue that there was a much earlier dispersal of modern humans from Africa, sometime before the eruption of Mount Toba in Sumatra (~74,000 years ago) [187, 188, 193].

The first model (the most widely accepted one), supports a single rapid wave of migration around 50,000 to 60,000, via southern route leading firstly, to the initial peopling of SEA and Sahul, and subsequent peopling of all of East Asia [194]. According to this model, the first settlers remained close to coastal habitats due to the abundance of marine food resources (fish, sea birds, sea mammals, etc.), water supplies, and raw materials for stone-tool manufacture [191, 195]. An increase of populations may have triggered the need for early humans to move from one coastal location to another towards east along the coastline, reaching Sahul 45,000 to 50,000 years ago.

One logical assumption for the time of expansion of the first modern humans, is the palaeoenvironmental context at that time, characterized by climate shifts between colder and warmer phases. The warmer and moister climate phase opened a green passage between Arabian Sea and the Zagros corridor, making it favourable for humans to expand from the arid deserts in Central Asia and northern Africa [100, 194]. There is also evidence of an inland savannah corridor throughout Sundaland around 45,000 to 60,000 years ago [175]. Another important aspect is that during the colder phases, because of the extended polar ice caps, the global sea level was considerably lower than it is today, enabling humans to easily cross water obstacles, or, even migrate via land bridges. For instance, around 65,000 and 70, 000 years ago, the widest strait in SEA that populations had to cross *en route* to Sahul (nowadays between Timor and Australia) was only 170 km [194].

Archaeological and palaeontological evidences have also been interpreted to support that initial settlement was under way by 50,000 to 60,000 years ago. Remains of the early pioneer settlement of SEA were found in Borneo and Australia, at the Niah Cave and Lake Mungo, respectively, dating more than 40,000 years ago [196, 197]. The same goes to the north, in East Asia, where remains of modern humans were found in the Tianyuan Cave dating 42,000– 39,000 years old [198]. More recently, remains of a human cranium were found in the Tam Pa Ling‘s cave, meaning Monkey Cave, in Laos, dating 46,000 to 63,000 years ago [199], and a small-bodied hominin found in the Callao Cave, provided the oldest evidence for the presence of early modern humans on the island of Luzon, as early as 67,000 years old [200].

2.2.2. Postglacial population movements

The climatic changes at the end of the Ice Age have been consider major triggers for prehistoric migrations of modern humans in SEA [165, 172]. During the last glaciation, the sea fell by up to 120 m below its present level, which opened up a vast low-lying area comprising MSEA and western ISEA (fig.12). However, the climate improvement at the end of the LGM, and subsequent rising of the Holocene Sea covering massive areas of lowland SEA, forced the hunter-gatherers living in the region to disperse and adapt to the new environmental conditions.

These disastrous events, most probably, had a catastrophic impact on the populations living in the region. Hunter-gatherers dispersals at the end of LGM have been associated with records of the Hoabinhian industry (that contain flaked, cobble artefacts), found in, mainly, south MSEA dating around 18,000 years ago, and in Malaysia and Sumatra at least 13,000 years ago. Interestingly, there is a gap in the archaeological assemblages in ISEA between 30,000 to 8000 years ago. The possible reason for this, is that, most likely, during that time populations were living on the coast near marine food resources. During the flood episodes, the coastal areas were drowned erasing the traces of human activity [163, 165].

2.2.3. Nusantao Maritime Trading Network hypothesis

The term “Nusantao” was proposed by Wilhelm Solheim [201] to designate the maritime-oriented natives (and their descendants) of Southeast Asia and coastal regions of China, Japan, Korea and Pacific. Solheim argues that the Nusantao people are associated with the spread and development of the Austronesian language, as the result of a very widespread complex trading and communication network within the last 10,000 years ago, covering the coastal areas of the Bay of Bengal and the Indian Ocean as far as Madagascar, and insular and continental SEA and the coastal areas of the China Sea and Japan. Solheim's concept of the Nusantao Maritime Trading and Communication Network (NMTCN) consists of Proto-Austronesian, Malayo-Polynesian and non-Austronesian-speaking seafaring populations continuously mixed genetically, culturally and linguistically

in a trade network, which was responsible for the cultural similarities in the Asia-Pacific region [176]. Solheim divided the NMTN into four geographic “lobes” of part of East Asia, SEA and Oceania (fig. 14). The four lobes are: the Northern Lobe, the Western Lobe, the Eastern Lobe and the Central Lobe.

The Northern Lobe comprises the trade network extending from northern SEA, including Taiwan and South China, to coastal Korea and eastern Japan, dating to at least 6,000 years ago [202]. The close prehistoric cultural relationship between populations of coastal Japan and Korea, and the people of SEA, has been supported by both physical anthropological, linguistic and archaeological evidences. The specific type of features shared in the dentition of all these populations (namely the Sundadont dentition pattern) [203, 204], and the several archaeological artefacts found widely dispersed throughout northern SEA (South China in particular), Korea and western Japan, suggesting a continuous network of population movements [202]. Similarly to these archaeological evidences, Solheim also suggested that some Austronesian elements, especially the ones related to rice agriculture vocabulary, were also brought north by Nusantao travellers, driven by the development of agriculture.

The Western Lobe of Solheim’s Nusantao concept represents the wide region extending from western Indonesia and Malaysia, throughout the coastal region of India and Sri Lanka to the west coast of Africa and Madagascar. Archaeological evidences suggest that this trading network started at least 4,500 years ago, based on specific earthenware vessels found throughout MSEA. The Eastern Lobe of the NMTN extends from the Moluccas throughout the Pacific, as far as Easter Island. This component is divided into two smaller lobes, an Early Eastern Lobe and a Late Eastern Lobe. The former ranges from the Moluccas in eastern Indonesia through the northwest region of Melanesia. The later, includes the Moluccas eastward to Wallacea and throughout the Pacific Ocean (except a large region of the coast and interior of New Guinea) [176]. The trading network probably started around 4,000 years ago. The spread of the Lapita cultural complex has been portrayed as a successful example of a NMTN in the Pacific. According to Solheim, this characteristic geometric dentate-stamped pottery seems to have appeared in Vietnam before 4,000 years ago, and later was brought to the Bismark archipelago by Nusantao traders [176].

At last, the Central Lobe is also further divided into two smaller lobes related to the phases of cultural spread: the Early Central Lobe - considered the homeland of the early NMTN - and the Late Central Lobe [176]. The Early Central Lobe was located in the eastern coast of Vietnam around 11,000 years ago, prior to the development of the

Austronesian language. These early Nusantao people spread to the Late Central Lobe around 7,000 years ago, where the Austronesian family language was developed. According to Solheim, the NMTN people speaking a Proto-Austronesian language spread both to Philippines via ISEA, and to Taiwan via South China. These last were isolated from the neighbouring populations, and consequently developed their distinct Austronesian languages. This linguistic process did not interfere with the development of the Malayo-Polynesian Austronesian language branch in ISEA.

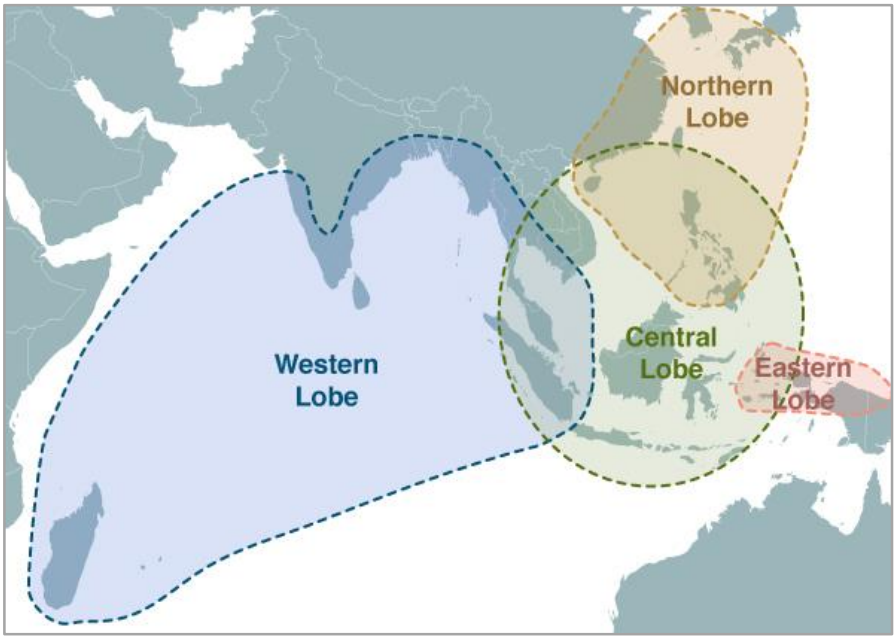


Figure 14. Map of the geographic division of the four lobes of Solheim's Nusantao Maritime Trading and Communication.

The technologically advanced Nusantao people, with large maritime capabilities, were able to adapt and incorporate new elements from the cultures they came in contact with, and developed a successful network of trade and communication; the Malayo-Polynesian language became part of the culture spread by the Nusantao traders in its expansions. Around 5,000 years ago the Nusantao people spread towards Micronesia in order to form the Early Eastern Lobe, and later spread west towards Malaysia continuing along the coast of India and Sri Lanka up to the western coast of Africa and Madagascar; and over time, further eastward towards into the Pacific as far as Easter Island [176]. Contrastingly, the opposite Bellwood's theory formulates that the Austronesian language spread from Taiwan to Philippines and from there eastward and westward to ISEA and Pacific. Aside from the origin of the Malayo-Polynesian language, the main difference between Bellwood's and Solheim's theories, is that the former suggests a linear

expansion, while the later suggests a reticulated network of expansions all overlapping the geographical area of the Late Central Lobe (which includes the Philippines).

2.2.4. The two-phased Neolithic dispersal hypothesis

Anderson's view of the linguistic and archaeological evidences related to the late Holocene colonization of ISEA, supports a process of "neolithicization" in the region, involving two phases of dispersal at different times and possibly of different character, proposed as Neolithic I and Neolithic II (fig. 15) [205]. The Neolithic I represents an earlier dispersal from the southwest, starting within the region of South China and then through Thailand and Vietnam, and reaching Borneo via Malaya, seen in basket or cord-marked ceramics dispersed throughout the region, dating at least 4,500 to 5,000 years ago. This expansion has been associated also with the dispersal of Austroasiatic languages [185, 206]. The Neolithic II refers to the hypothetical "out-of-Taiwan" migration, associated with the dispersal of Austronesian languages [178]. It led to the expansion of farming and the characteristically red-slipped pottery in the Philippines and elsewhere in the eastern ISEA, and later into Malaysia and Vietnam [205, 207]. The widely accepted out-of-Taiwan model is going to be discussed in more detail in the next section.

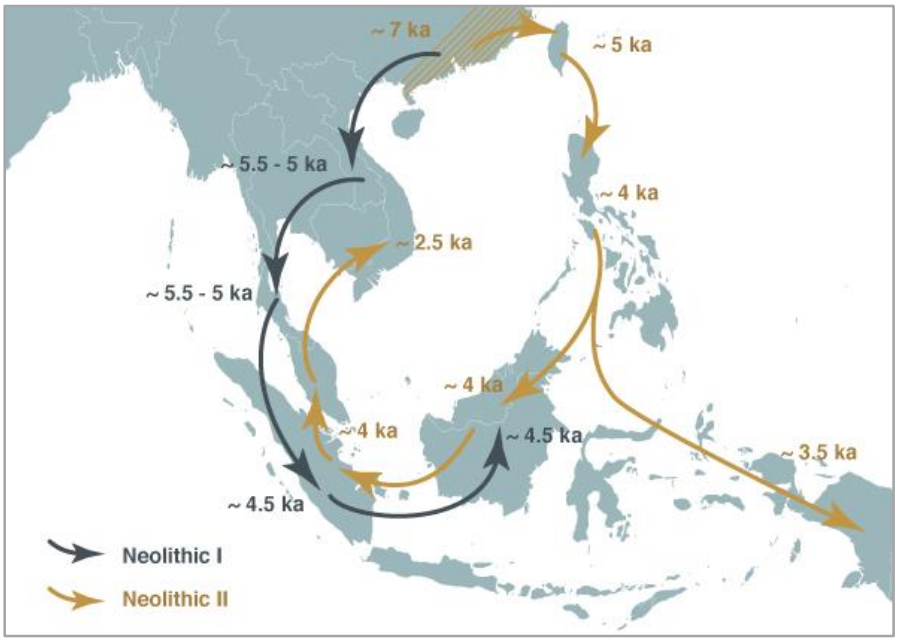


Figure 15. Schematic representation of the Neolithic I and II reticular pattern. Adapted from [205].

2.2.5. The Neolithic farming/language dispersal hypothesis: the Austronesian dispersal in ISEA

Over the last decades, the theories/hypotheses concerning the colonization of ISEA, have been dominated by variants of the Austronesian dispersal hypothesis [185, 186, 207-210]. This theory, along with other models for the spread of certain languages, such as the Bantu across sub-Saharan Africa [211, 212] or the Indo-European across Eurasia [213], represents an example of a proposed global phenomenon so called the farming/language dispersal hypothesis [208, 214]. According to this hypothesis, the main factor behind the spread of many of the major language families was the development of agriculture. The subsequent increase of population density, led to demic expansions of farming populations along with their language and culture [185].

In the case of the Austronesian language hypothesis, also called, the “out-of-Taiwan” (OOT) model, Neolithic farmers dispersed from mainland South China to Taiwan and thence to ISEA and the Pacific, carrying with them a new material culture and the precursor of the Austronesian languages that are spoken in the region nowadays [185, 215, 216]. The OOT cultural package included agricultural technologies, such as domestic rice and animals (pigs, chickens and dogs), and other Neolithic traits, such as pottery and polished stone adzes [185, 208]. It has been proposed that these advanced technologies would probably enabled the Austronesians speakers to colonize and, to some extent, replace pre-existing groups of hunter–gatherers occupying ISEA [185, 216].

2.2.5.1. Linguistic evidences of the OOT

As previously mentioned, the Austronesian language family comprises the languages most widely spoken throughout ISEA. The OOT model mainly relies in the linguistic differentiation within the Austronesian language family tree, to establish the Neolithic population movements in ISEA and Pacific [217] (fig. 16). The standard structure of the Austronesian language family tree is very hierarchical, with bifurcations corresponding to inferred movements from Taiwan, where it is found at higher diversity (the Proto-Austronesian homeland, where nine of the 10 putative primary branches are found) through western ISEA (where several languages of the Western Malayo-

Polynesian branch are spoken) and eastern Indonesia (where a sub-branch emerged, the Central Malayo-Polynesian). The migration continued east across northern New Guinea (where languages of the Eastern Malayo-Polynesian branch emerged), and finally into the Pacific (evolving into the Oceanic languages, including Polynesian and Micronesian) [217].

The direction and structure of the Austronesian language tree, with all the primary branches, but one, spoken only amongst Taiwanese aborigines, and the tenth Malayo-Polynesian branch giving rise to all of the remaining non-Formosan Austronesian languages spoken throughout the region, places Taiwan as the point of origin for the Austronesian language family [178, 209, 216-218].

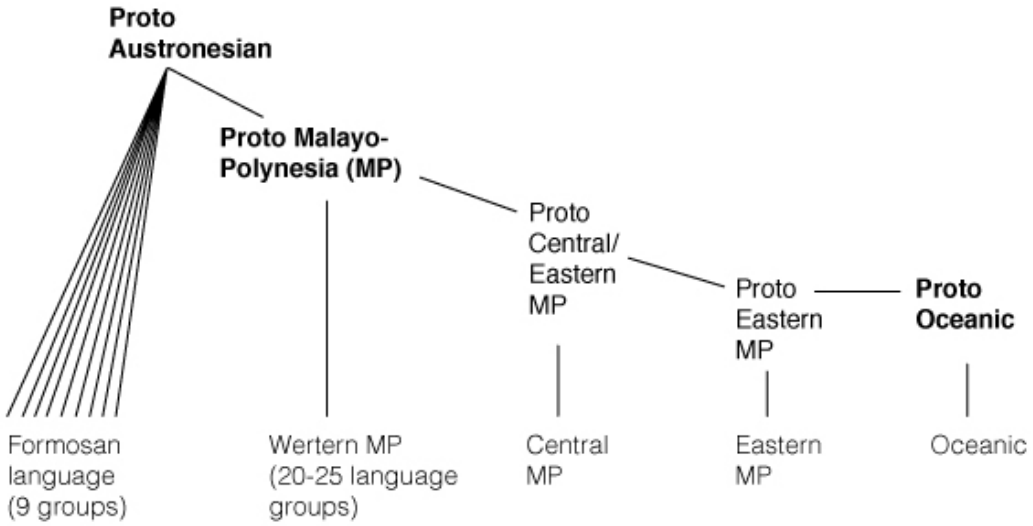


Figure 16. Structure of the Austronesian language family tree phylogeny according to Blust (1995) [217].

Recent studies have shown that the Austronesian language family structure is not as hierarchical as previously assumed; it is rather flatter at multiple levels [214, 218, 219]. Blust's reconstruction of the Austronesian language tree relied on shared innovations in inherited linguistic features to define phylogenetic subgroups. The number of shared innovations validates the Malayo-Polynesian branch as subgroup, but the same does not apply to the Western Malayo-Polynesian branch, because though its languages show the Malayo-Polynesian innovations, they do not show unique innovations relative to Malayo-Polynesian languages assigned to the Central/Eastern Malayo-Polynesian branch [218]. Therefore, in the new phylogeny of the Austronesian language tree, the Western Malayo-

Polynesian branch ceases to exist as a separated group; instead several individual branches radiate from Proto-Malayo-Polynesian branch, including the Eastern Malayo-Polynesian branch (Central Malayo-Polynesian node also ceases to exist) (fig. 17).

The Proto-Oceanic branch also suffered some alterations: instead of showing the hierarchical structure expected under the OOT model, the Oceanic language structure is flat and rake-like, suggesting a multi-directional expansion with point of origin in the Bismarck islands [220, 221].

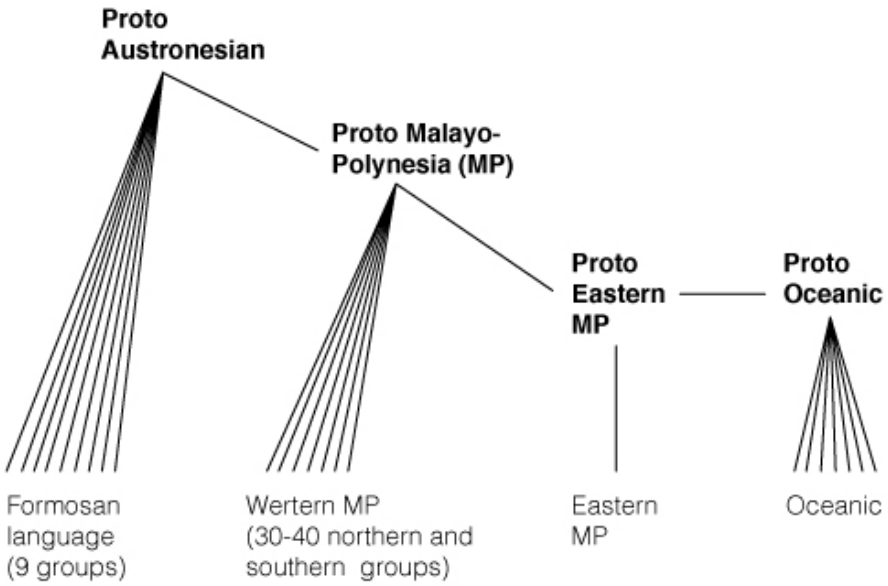


Figure 17. Structure of the Austronesian language family tree phylogeny according to the most recent studies [218, 219].

The Bayesian phylogenetic analysis of Malayo-Polynesian lexical data supported the OOT model, dating the origin of proto-Austronesian approximately at 5230 years ago in Taiwan, and the development of the Malayo-Polynesian branch somewhat within the last 4,500 to 3,000 years ago [215]. Bellwood also dated the branches in the Austronesian language tree by correlating them with the split of Austronesian-speaking populations. Based on this dating method, the Proto-Austronesian dates around 5,500 years and the Austronesian around 4,500 years [163].

2.2.5.2. Archaeological and archaeobotanic evidences of the OOT

Archaeological records have been interpreted to chronologically reconstruct the direction and the cultural components of the Austronesian expansion. Assemblages containing the typical Neolithic Austronesian speakers cultural package have been found throughout South China, Taiwan, ISEA and the Western Pacific dating 6,000 to 3,500 years ago. Moreover, the estimated ages of the archaeological records evidence a north-to-south geographic distribution pattern, in which the excavated sites with the earliest findings were found in China, Taiwan and Luzon, and progressively later, in Indonesia and Pacific (Fig. 18) [222-224]. Therefore, the evidences of cultural material, specially pottery, have been used as chronological markers for the sequential Neolithic migration starting around 6,000 years ago from South China to Taiwan, later on to Philippines around 4,500-4,000 years ago, to eastern Indonesia around 4,000–3,500 years ago, and thence across the north of New Guinea (largely excluding mainland New Guinea) to the Bismarck by ~3,500–3,300 years ago, and to the Melanesian islands of Solomon and Vanuatu by around 3,000 years ago [222-226].

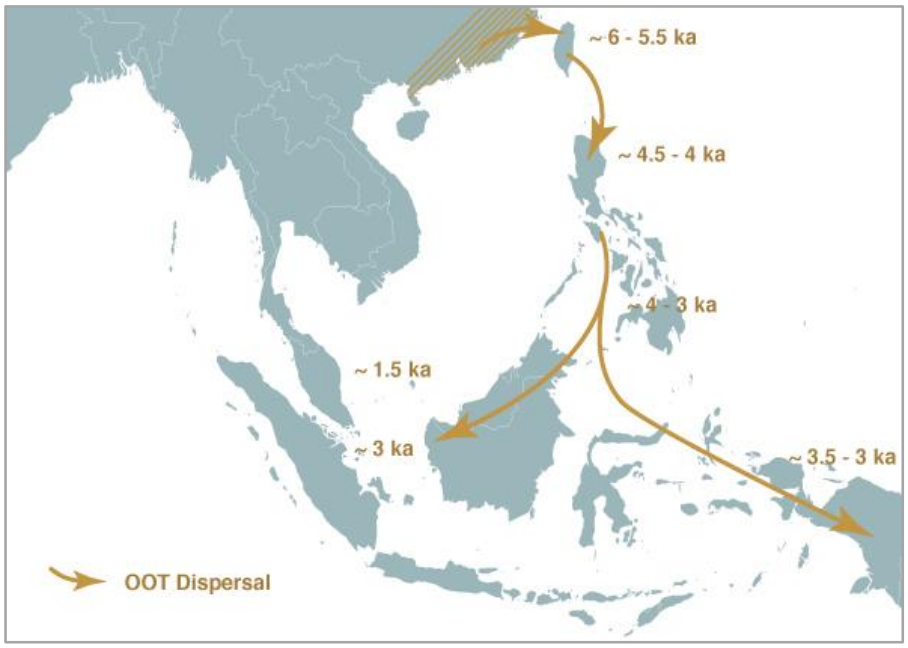


Figure 18. Summary of the Austronesian dispersal model (out-of-Taiwan hypothesis).

The earliest archaeological evidences of the Neolithic culture in Taiwan dates to around 6,000-5,500 years ago, presumably carried by Pre-Austronesian speaker farmers moving from mainland China across the Formosa Strait. The archaeological assemblages

of Taiwanese Tapenkeng culture, characterised by cord-marked globular pots with carved everted rims and polished stone adzes, resemble those from excavated sites in southeast coast of mainland China (Fujian and Guangdong), comprising potsherds decorated with carved lines, impressed semicircles and stamped dentate patterns inside incisions [207]. Evidences of agriculture in Taiwan, based on rice and pollen vestiges, dated to 5,000 years ago. This time frame also fits the development of the Proto-Austronesian language [209].

The Taiwanese assemblage most similar to the early Philippines and Indonesia pottery belongs to the Yuanshan culture, one of the regional complexes that differentiated from the Tapenkeng culture between 4,500 and 3,000 years ago. The Yuanshan pottery consists in globular vessels, some of them with strap handles and ring feet. In this culture, the pottery is, usually, painted with reddish brown pigment either incised or punctuated/stamped, and there is no evidence of the initial cordmarking [227]. Based on these unique characteristics, that are also found in the earliest pottery assemblages in Philippines and Indonesia, the Yuanshan culture has been proposed by Bellwood (and widely accepted by the archaeological community), as the immediate ancestor culture of the Neolithic cultures in ISEA [227, 228].

Archaeological assemblages containing plain or red-slipped pottery similar to the one of the Yuanshan culture, as well as imprints of rice shells, found in several excavated sites throughout Philippines and ISEA, dating to 4,500 – 3,500 years ago, suggest that the agriculture was underway ISEA, from Taiwan, by at least 4,000 years ago, and probably earlier [229, 230]. During that time frame, the Proto-Austronesian language differentiated into the Formosan branches (spoken only in Taiwan) and the Proto-Malayo-Polynesian branch as populations dispersed to Philippines, and later on divided into several branches of Western and Eastern Malayo-Polynesian, as Neolithic farmers reached the Indonesian archipelago [209].

By 3,500 years ago, agricultural Austronesian speakers took two different routes, one heading west towards Java, and reaching certain regions of MSEA (in Vietnam and in the Malaysian Peninsula), by 1,500 years ago. Vestiges of the Neolithic quadrangular and pick adzes, that are related with early red-slipped pottery sites in the Philippines, were found in East Java [163]. During that time, other Austronesian-speaking populations headed east towards the Pacific, settling as far as Easter Island, taking with them the precursor of the Lapita pottery present in Melanesia by 3,500 years ago [231]. The development of the Lapita culture was marked by the gradual disappearance of red-slipping in the pottery assemblages. Instead the Lapita earthenware pottery was coiled or

slab-built vessels often decorated with a complex range of dentate-stamped motifs. There was also a shift from rice cultivation towards other cereal and tuber crops [231]. Lapita assemblages, containing not only the distinctive geometric dentate-stamped pottery, but also other artefacts, such as ground-stone adzes and obsidian flaked-stone tools, were found in coastal excavated sites from the Admiralties to Samoa, suggesting that this cultural technocomplex arose, most probably, somewhere within the Bismarck Archipelago, and scattered over a large area into to previously uninhabited Pacific Islands, about 3,500 – 3,000 years ago [216, 222, 231, 232].

2.2.5.3. The inconsistencies/ limitations in the integrated perspective of the OOT

The Austronesian dispersal has been proposed as an archetype of the farming/language dispersal hypothesis, with the associated linguistic replacement, genetic clines and Neolithic packages. The theory suggests that rice agriculturalists dispersed from Taiwan into the Philippines and Indo-Malaysia and, after largely bypassing New Guinea, moved into the Pacific, carrying with them the Austronesian languages [185, 208]. This dispersal is presumed to have been fuelled by the development of agricultural technologies, primarily based on rice but including other crops and some domesticated animals (pigs, dogs and chickens), and other Neolithic items, including pottery and polished stone adzes [178, 185, 207, 208, 216]. Noteworthy, none of these material-culture types are necessary or unique markers of farming in ISEA. Although archaeological evidence supports a link between Austronesian origins and a farming capacity in the region of the Taiwan Strait, evidences of a linguistic expansion southwards involving the transplantation of the same subsistence practices to ISEA is lacking. In fact, the Austronesian populations instead of continuing the cultivation of essentially the same crops as had been domesticated in the beginning of their Neolithic journey, shifted their subsistence economy from grain cultivation to the root and arboreal crops in most of ISEA and completely in the Pacific. Taking this into consideration and the geographic localization of the Austronesian speakers (mostly in islands), it seems more plausible that the enabling technology for this dispersal was the advanced maritime capacity to explore the marine resources and of locally-adapted crops in the region [233]. The assumption

that the southward expansion was predicated on maritime foraging and trade has been used to support other alternative models [100, 176, 202, 234].

The proponents of the OOT model draw it essentially from historical linguistics and some archaeological evidences. However, evidences supporting a dispersal of farming practices with early Malayo-Polynesian-speaking voyagers across ISEA, during the initial period of the Austronesian expansion, are very limited. There is an almost complete lack of domesticated rice, or the associated farming technologies and food processing and consumption, dating to 3,000 - 4,000 years ago. From all ISEA, only two sites presented both pottery and rice remains, that may date to the Neolithic arrival, one at Gua Sireh in western Sarawak [235] and the other at Ulu Leang in southern Sulawesi [236]. Furthermore, it has been suggested that one of these sites, the Gua Sireh (due to its location and age) may represent a Neolithic arriving from mainland Asia rather than Taiwan [235]. In fact, the earliest records of pottery across ISEA suggest the occurrence of multiple incursions into the region by at least two different cultures, one from MSEA carrying a corded or basketry-wrapped paddle impressed pottery, and a second influence of the red-slipped pottery tradition originated in Taiwan, and spread by the supposed OOT dispersal [214, 222].

Other elements claimed as components of a Taiwanese cultural package, such as some domestic animals (dogs or domestic pigs), also provide insufficient support for the Taiwan-based dispersal of farmers into ISEA. Firstly, the northernmost evidence of an early faunal introduction is in the Niah Cave in Sarawak, but there are some uncertainties regarding the dating [237]. Additionally, recent genetic studies revealed a clear association between pig populations (*Sus* spp.) from MSEA, ISEA, New Guinea, and Oceania, suggesting that they were introduced into New Guinea, via eastern ISEA, not from Taiwan, but instead from MSEA (region proposed as the centre of pig and chicken domestication [214, 238]) [239, 240]. Additionally, other technologies associated with the spread of Austronesian populations, such fishhooks, shell beads, and shell adzes have been found in ISEA and Melanesia, predating the time of arrival of the Neolithic pottery in those regions [198].

Overall, both linguistic and archaeological evidences, or even plant and animal domestication histories, suggest that most of the components that have been presumed to have diffused with the Austronesian language from Taiwan were already present in ISEA. Nevertheless, some archaeological records, such as the red-slipped pottery are still considered a reliable chronological marker for the arrival of the Austronesian speakers in the region [222, 223]. Taking into consideration the complex records of the human

footprint in ISEA, it becomes clear that other major forces, aside from language and agriculture, shaped the demographic history of ISEA.

3. MAIN GENETIC EVIDENCES OF POPULATION MOVEMENTS IN ISEA

3.1. UNIPARENTAL GENETIC MARKERS

Over the last decades, genetic research has been an invaluable tool for understanding the human origins and migrations in the world. Until recently, the genetic information (in particular from uniparental genetic markers, mtDNA and Y chromosome) of the region of SEA have been analysed in light of the two widely accepted major demographics events in the region: the initial settlement by anatomically modern humans in the Late Pleistocene, followed by the more recent Holocene dispersal of the populations speaking Austronesian languages.

It is relatively well establish that the first settlers of the ISEA migrated from Sunda land bridge that connected MSEA to regions as far east as Wallace's Line approximately 50,000 years ago [196, 241, 242]. The genetic signatures of the initial colonization of Sundaland and Sahul are in agreement with the archaeological picture of an initial colonization by modern humans at least 50,000 years ago. Several mtDNA lineages, with a patterned distribution across both present-day MSEA and ISEA, trace back to the main branches of macrohaplogroups M and N in that time frame [104, 154, 172, 243-246]. For instance, haplogroup M9, present in SEA, Taiwan and China, most likely arose in SEA approximately 50,000 years ago, and its major subclade, haplogroup E, around 30,000 years ago [172].

The second most widely accepted major demographic event in the region is associated with massive agriculturally-driven migrations in the mid-Holocene (OOT). According to this model, the largest ancestral contribution to ISEA populations trace back to Taiwanese Neolithic populations. mtDNA has been extensively used to argue for or against the OOT model of the Austronesian language expansion around ~4,500 years ago. The support for the Taiwanese origin of this model relies mostly on the geographic

distribution of a specific D-loop motif (variation at nucleotide positions 16217, 16247 and 16261) with a nine base pair deletion in the COII/ tRNALys intergenic region, which is characteristic to the haplogroup B4a1a1a lineages of all Austronesian speaking populations. Since this haplogroup is present at high frequency throughout the Pacific, especially in Polynesia, it is referred as the Polynesian motif (PM) [97, 247, 248]. The ancestral haplogroup of the PM, B4a1a, with an Asian origin, is found in Taiwan and throughout ISEA and the western Pacific. The fact that the PM could be traced back to Asian populations (in particular Taiwan) was seen initially as supporting an OOT for the Polynesian origins [249, 250].

However, these initial assumptions relied on the analysis of poorly resolved mtDNA control-region sequences. Recent analyses of complete mtDNA genomes from haplogroup B4a1a provided new insights about the immediate origin and age of B4a1a1a [248]. B4a1a1a is not found in Taiwan or the Philippines, and based on the estimates of coalescence dates, it most likely originated within the Bismarck Archipelago around 8,000 years ago, several thousand years before the Lapita cultural complex (associated with the Austronesian language expansion). In fact, the expansion of this lineage most likely resulted from the Early to Mid-Holocene sea level rise rather than from the dispersal of Austronesian-speaking agriculturalists [248, 251, 252].

With a similar expansion pattern to haplogroup B4a1, haplogroup E lineages also appear to have been carried with the postglacial population movements in ISEA [172]. Haplogroup E arose most likely in northeast Sundaland around 35,000 years ago and was caught up in the dramatic episodes of population dispersals that began in eastern Sundaland/northwest Wallacea about 12,000 years ago [172], reaching Taiwan somewhere between 8-4 ka [172] (according to the recalibration of the mtDNA molecular clock [91]). The dispersal of this haplogroup most likely reflects the impact of the climate changes in the Early to Mid-Holocene, that triggered multiple population movements (possibly hunter gatherers who followed the now-submerged river systems in the region) [253], rather to what has been suggested that haplogroup E may be associated with early Austronesians and the subsequent dispersal of Austronesian languages [254]. This last assumption was based on the analysis of a complete mitochondrial DNA genome sequence of an ~8,000-year-old skeleton from Liang Island (located between China and Taiwan), supporting a link between southern China and Formosan populations. These finds not necessarily mean that the most ancestral E lineages arose between China and Taiwan (as postulated by the authors [254]), in fact, the overall pattern of haplogroup E strongly suggests ISEA as the point of origin and evolution for the clade. Therefore the

loss of many lineages during the postglacial period in ISEA may potentially explain the lack of ancestral lineages, such as the ancient mtDNA E1 lineage detected further north [254].

Haplogroup E lineages suggest that Taiwan may be the recipient of population movements from the south, before the Austronesian dispersal, rather than being the major source of Holocene population expansions across ISEA [172, 251]. In fact, an early extensive study of the mitochondrial gene pool of ISEA had suggested that only a maximum of 20% of present-day lineages could have resulted from the mid-Holocene Austronesian language dispersal, and the majority of the ISEA lineages dates to Late Pleistocene and Early Holocene [243]. For the Y chromosome, most of the ISEA lineages resulted from pre-Neolithic demographic expansions (most likely carried by the first settlers and subsequent postglacial expansions), and only a small fraction of lineages are thought to have arrived with the Austronesian-speaking populations from Taiwan [128, 255-258]. Karafet et al.[255] argued for a discontinuous four-phase colonization process with several population incursions in SEA. According to this model, the initial settlers introduced the basal Y haplogroups (C and K lineages) in the entire region. In the second phase, the hunter-gatherers in the late-Pleistocene/postglacial migrations introduced several major subclades of haplogroup O in ISEA from MSEA. Later, a small fraction of haplogroup O subclades entered ISEA in the mid-Holocene from Taiwan with the Austronesian expansion. Last stage of this model refers to the population moments in the historic era, from southern Asia, Arabia and China, which led to dramatic cultural and social changes, but had small impact in the genetic composition of ISEA.

This complex pattern of settlement and dispersals in SEA was corroborated in recent studies [252, 256]. Trejaut et al. [256], by analysing the male genetic diversity of SEA populations, proposed a pincer model that includes northeastward dispersal from Taiwan into Philippines and Indonesia and a southward dispersal via the Indochinese peninsula to Indonesia and into the Philippines and Taiwan, within the last 20,000 years ago. This time frame reinforces previous mtDNA and Y chromosome studies that concluded that the major genetic contribution for present-day ISEA gene pool relates to population moments in the late-Pleistocene to early-Holocene [172, 255].

Interestingly, the male genetic diversity in the Pacific does not trace back to ISEA Austronesian speaking populations like in the mtDNA analysis. Austronesian-speaking populations in both Near and Remote Oceania (with Asian derived mtDNA lineages) carry Y chromosomes lineages belonging to haplogroups K, M, S and C, which are restricted to Oceanic populations and are thought to have originated in Near Oceania during the

Pleistocene period of human occupation [127, 259, 260]. It has been suggested that the genetic composition of the Oceanic populations resulted from the intermarriage between Austronesian-speaking females carrying Asian mtDNA lineages (for example: mtDNA PM) with male Melanesians *en route* to the Pacific, suggesting that the Lapita societies had a matrilineal structure and a matrilocality residence pattern [247, 261]. These conclusions were based on the assumption that the PM was a genetic marker of the OOT, however according to Soares et al. [248], although the motif (or its ancestor) has an ultimately mainland Asian ancestry, the populations carrying the PM arrived in Near Oceania ~6–10 ka, thus they were already well-established in Near Oceania by the time of the arrival of the so called Neolithic “out-of-Taiwan” populations. Therefore, it appears that the Neolithic OOT migration did not leave a substantial signal, both on the mtDNA and Y-chromosome gene pool in the ISEA and Remote Pacific Islanders, in the last 3,000 years.

3.2. AUTOSOMAL GENETIC MARKERS

Over the last years, genome-wide studies have been employed to overcome some of the limitations of mtDNA and Y chromosome analyses. However, there is still a relatively limited number of studies using autosomal genetic markers in the region of SEA, in comparison to analyses using the uniparental genetic markers.

The first studies using autosomal STRs focused on Pacific populations, and provided a similar picture of the population prehistory shown by mtDNA and Y chromosome studies [154, 260]. These two studies showed that Austronesian-speaking Pacific islanders have a strong East Asian ancestry (almost 80%) with a smaller Melanesian contribution (~20%), and only along the coast and through the islands of the Bismarck Archipelago, suggesting that they were almost entirely the result of the Austronesian expansion around 4,000 years ago from Taiwan towards New Guinea.

With the development of high-throughput SNP genotyping platforms, more comprehensive genetic studies were employed. Analyses using the HUGO Pan-Asian SNP dataset, also indicate an East Asian derived population migration towards New Guinea, with the admixture in the time frame of the Austronesian expansion [262, 263].

All the results mentioned above should be interpreted with caution, since in the first two studies case, few samples from ISEA were included, and in the third more recent

study, samples from Taiwan were not included. This fact is particularly important, since more comprehensive genome-wide analyses revealed a picture of prevalent gene flow among Asian populations, in which aboriginal Taiwanese appears to be an offshoot from Island Southeast Asians [157].

Recently, Jinan et al. [264] based on genome-wide analyses suggested a more complex migration history than that of the two-wave hypothesis generally accepted for the peopling of ISEA. This study argued that SEA populations are divided into two clusters with clear different genetic composition. The “island” cluster includes the Taiwanese, Filipino, and Sulawesi, whereas the second cluster, referred as the “mainland” cluster, is represented by MSEA populations and the Austronesian-speaking populations that were previously part of the Sunda landmass (Malaysia and Java and Borneo islands). The western ISEA ancestry to mainland sources is in agreement with the genetic signals of the late-Pleistocene/early-Holocene migrations from Indochina southward towards Malaysia, Sumatra, Java, and Borneo, that were previously shown in the mtDNA and Y chromosome analyses [172, 255]. According to the authors, this admixture pattern suggests that Taiwanese populations were not the sole contributors to the genetic diversity in all Austronesian groups [264].

Genome-wide studies have been interpreted as supporting both Taiwan-derived [154, 260, 263] and multiple-wave [264] migration models for the dispersal of Austronesian-speaking populations. However, these studies did not allow to establish, definitely, the point of origin of the Austronesian language. Two recent studies addressed that question. Lispson et al. [265] using a relatively extensive genome-wide dataset comprising 31 Austronesian-speaking populations and other 25 populations from the HUGO Pan-Asian SNP Consortium [157] and the HGDP [161], concluded that the Austronesian ancestry component is most closely related to present-day Formosans. According to this study, the Asian ancestry component found in the western ISEA populations could be, most likely, explained because their ancestral populations spoke Austroasiatic languages. The Taiwanese ancestry of the Austronesian genetic component was also suggested in another recent autosomal study [266], that further proposed the southern Chinese Daic domain as the cradle of Proto-Austronesian migrants to Taiwan. However, caution is needed when analysing their conclusions, due to their limited number of SNPs. In particular, in spite of the relatively large study performed by Lipson et al. [265], these authors only used a very limited number of SNPs (not even reaching 10,000 SNPs), which weakens their conclusions that the Austronesian expansion involved substantial human migrations.

Indeed, so far, genome-wide studies are still lacking in the region of SEA. There is a clear necessity to perform more comprehensive studies, with wider sampling, not only from SEA populations but also from neighbouring regions, and with larger sample sizes and more SNPs.

4. BIOMEDICAL CONTRIBUTIONS OF EVOLUTIONARY POPULATION STUDIES

In the last decades, the advances in molecular genetic technology has led to a new era of human genetic studies. As a result, large amounts of genomic data in human population have become available. The study of this genetic variation has both evolutionary significance and biomedical applications. It can help to understand ancient human population migrations and interactions, as well as the resultant genetic structure and differentiation between, and within, different human population groups. The study of human genetic variation may provide invaluable information for biomedical approaches, since the human population is not homogeneous in terms of risk of disease [267]; for instances some disease-causing alleles occur more often in people from specific geographic regions [268-270].

Additionally, recently analyses of thousands of markers on a genome-wide scale (genome-wide association studies - GWAS) have revealed numerous disease-associated loci and have provided significant insights into the allelic architecture of complex diseases traits [269, 271]. Several studies have started to use this approach to explore the interaction between population evolutionary history and complex disease traits, such as type 2 diabetes variants [272] and hypertension [273]. However, in the case of admixed populations, the genetic variation can produce spurious association of genotypes and phenotypes through their separate associations with ancestry. In these cases, knowledge of the population genetic structure becomes essential to reduce the false-positives [271]. Therefore, population-based genetic studies, by increasing the knowledge of the genetic characterization of human populations, can enhance the understanding of the genetic basis of complex diseases, and provide the foundation for the development of numerous multidisciplinary researches.

CHAPTER 2 - AIMS

The complex demographic history of SEA, tangled with multiple Holocene migrations and population displacements following the first Late Pleistocene settlement around 50-60ka, has led to the establishment of one of the most genetically diverse population in the world. Understanding of how genetic diversity is structured in a population as well as the forces that shaped that diversity, is important not only for anthropological and evolutionary fields, but also for biomedical research. For example, knowledge of the population structure is crucial to minimize bias in clinical research, since individuals from different populations often respond differently to medical treatments.

In this sense, the main goal of this study was to contribute for a more comprehensive view of the human genetic variation of SEA, as well as to provide a better picture of the main dispersal routes and the impact of those dispersals on the population history in the region. Additionally, this study also aims to provide a genetic background for SEA populations, that could be used in future clinical and medical research. To achieve this, the research work was divided in three main specific goals:

1. **Perform a comprehensive study of the region combining the three genetic systems, mtDNA, Y chromosome and genome-wide data**. To achieve this, a large dataset of new mtDNA (both control-region and whole-genome sequences) and Y-chromosome data was analysed, and a set of lineages were appointed as markers for postglacial and Neolithic movements. To test these two very different hypothesis for the prehistory of SEA (postglacial and OOT expansions), the phylogenies of the major mtDNA OOT candidates, M7, B4a1a and E, were reconstructed. Additionally, genome-wide patterns of a total of 23,332 SNPs from the Pan-Asian SNP Genotyping Database was compared to the mtDNA and Y chromosome results. These results are presented in the Paper I and constitute the groundwork of this thesis.

2. **Characterize at high resolution selected maternal lineages that have been tentatively associated with various demographic events in SEA**.
   The sequence variation of whole-mtDNA genomes belonging to the remaining low frequency mtDNA lineages, previously identified by the founder analysis in Paper I, was studied in detail. The working hypothesis was to test mtDNA lineages associated with the first settlement (haplogroup F3, R9b), postglacial expansions (haplogroup R9c, N9a) and mid-Holocene dispersals from Taiwan

(haplogroups B4c1, F1a4, B5b, Y2, B4b1 and D5). The phylogeographic analysis was accomplished by using a comprehensive dataset of 114 newly sequenced whole-mtDNA genomes and 829 published whole-mtDNA genomes from a vast geographic region including Taiwan, MSEA, ISEA and Near Oceania. These results are presented in the Paper II.

3. **Evaluate the genetic relationships between Southeast Asian populations and provide a comprehensive ancestry landscape of SEA population history, using genome-wide SNPs variation.** The characterization of the genetic makeup of the region was accomplished by performing an in-depth population genetic study using genome-wide SNP data from 47 East/Southeast Asian populations (both new populations genotyped with the Illumina HumanOmniExpress BeadChip containing approximately 700,000 SNPs, and published populations in the 1000 Genomes database and the Human Genetic Diversity Project database). These results are presented in the Paper III.

CHAPTER 3 - RESEARCH WORK

STUDY OF MATERNAL LINEAGES IN SOUTHEAST ASIA

Paper I

Resolving the ancestry of Austronesian-speaking populations

Paper II

Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia

PAPER I

Resolving the ancestry of Austronesian-speaking populations

Soares P, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, Mormina M, **Brandão A**, Fraser RM, Wang T-Y, Loo J-H, Snell C, Ko T-M, Amorim A, Pala M, Macaulay V, Bulbeck D, Wilson JF, Gusmão L, Pereira L, Oppenheimer S, Lin M, Richards MB (2016). Hum Genet, 135(3), 309-326. doi: 10.1007/s00439-015-1620-z

This paper establishes the groundwork of the present research thesis, and I was involved in the mtDNA laboratory work and some statistical analyses (namely, interpolation maps of spatial frequencies of HVS-I sequences).

Hum Genet (2016) 135:309–326
DOI 10.1007/s00439-015-1620-z

ORIGINAL INVESTIGATION

Resolving the ancestry of Austronesian-speaking populations

Pedro A. Soares[1,2,3] · Jean A. Trejaut[4] · Teresa Rito[2,5,6] · Bruno Cavadas[2,7] · Catherine Hill[3] · Ken Khong Eng[3,8] · Maru Mormina[3,9] · Andreia Brandão[2,7,10,11] · Ross M. Fraser[12,13] · Tse-Yi Wang[4] · Jun-Hun Loo[4] · Christopher Snell[3] · Tsang-Ming Ko[14] · António Amorim[2,7,15] · Maria Pala[10] · Vincent Macaulay[16] · David Bulbeck[17] · James F. Wilson[12,18] · Leonor Gusmão[2,19] · Luísa Pereira[2,7,20] · Stephen Oppenheimer[21] · Marie Lin[4] · Martin B. Richards[3,10]

Received: 19 September 2015 / Accepted: 18 November 2015 / Published online: 18 January 2016

Abstract There are two very different interpretations of the prehistory of Island Southeast Asia (ISEA), with genetic evidence invoked in support of both. The "out-of-Taiwan" model proposes a major Late Holocene expansion of Neolithic Austronesian speakers from Taiwan. An alternative, proposing that Late Glacial/postglacial sea-level rises triggered largely autochthonous dispersals, accounts for some otherwise enigmatic genetic patterns, but fails to explain the Austronesian language dispersal. Combining mitochondrial DNA (mtDNA), Y-chromosome and genome-wide data, we performed the most comprehensive analysis of the region to date, obtaining highly consistent results across all three systems and allowing us to reconcile the models. We infer a primarily common ancestry for Taiwan/ISEA populations established before the Neolithic, but also detected clear signals of two minor Late Holocene migrations, probably representing Neolithic input from both Mainland Southeast Asia and South China, via Taiwan. This latter may therefore have mediated the Austronesian language dispersal, implying small-scale migration and language shift rather than large-scale expansion.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-015-1620-z) contains supplementary material, which is available to authorized users.

✉ Martin B. Richards
m.b.richards@hud.ac.uk

1 Department of Biology, CBMA (Centre of Molecular and Environmental Biology), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

2 IPATIMUP (Institute of Molecular Pathology and Immunology of the University of Porto), Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

3 Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK

4 Molecular Anthropology and Transfusion Medicine Research Laboratory, Mackay Memorial Hospital, Taipei City 10449, Taiwan

5 Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

6 ICVS/3B's-PT Government Associate Laboratory, Braga, Guimarães, Portugal

7 I3S - Institute for Research Innovation in Health, University of Porto, 4200-135 Porto, Portugal

8 Centre for Global Archaeological Research, Universiti Sains Malaysia (USM), 11800 Penang, Malaysia

9 Department of Applied Social Studies, University of Winchester, Sparkford Road, Winchester SO22 4NR, UK

10 Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK

11 ICBAS - Institute of Biomedical Sciences Abel Salazar, University of Porto, 4050-313 Porto, Portugal

12 Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, Scotland, UK

13 Synpromics Ltd, Nine Edinburgh Bioquarter, Edinburgh EH16 4UX, UK

14 Department of Obstetrics and Gynecology, National Taiwan University, Roosevelt Rd., Taipei 10617, Taiwan

15 Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

16 Department of Statistics, University of Glasgow, 15 University Gardens, Glasgow G12 8QQ, UK

Introduction

Austronesian languages are spoken throughout Taiwan, Island Southeast Asia (ISEA), parts of New Guinea and most of the Pacific Islands. The high linguistic diversity observed in the aboriginal groups of Taiwan, compared to the single language branch (Malayo-Polynesian) spoken throughout the remainder of this vast distribution (Ross 2005), has suggested to historical linguists a homeland on the island of Taiwan (Blust 1976, 1995). The Taiwanese linguistic homeland model has received further support in recent years from the work of Ross (2009), which effectively nests Malayo-Polynesian within the Formosan language tree, and has led in turn to the prevailing "out-of-Taiwan" model (Bellwood 1997; Hung et al. 2011) for the spread by demic diffusion of farming—referring specifically to rice agriculture—and red-slipped pottery from 4000 years ago (4 ka) in ISEA, culminating in the spread of Austronesian Oceanic speakers into the Pacific within the last 3 ka (Bellwood 1997; Spriggs 2003, 2007). In practice, the Neolithic in ISEA is defined by the appearance of ceramics, and less ubiquitously new shell artefacts and cloth and barkcloth technologies, and any role of rice or other introduced agriculture has proved much more contentious (Spriggs 2011).

The "out-of-Taiwan" model has been dominant for 30 years, despite challenges on many fronts. Languages can be transmitted horizontally, so a root in Taiwan need not automatically imply a demic diffusion model. More complex pictures are emerging for ISEA, in which coastal language shift and language transmissions play the major role (Donohue and Denham 2010, 2015). This paves the way for a modified small-scale "out-of-Taiwan" model that retains the linguistic argument for the origin of the Austronesian family in Taiwan without assuming any large-scale population movement or replacement, or that rice agriculture is the driving force (Diamond and Bellwood 2003; Oppenheimer 2004; Spriggs 2011).

In fact, a key driver of human mobility may have been the dramatic transformation in the landscape of ISEA in the late Pleistocene/early Holocene. Sea-level rises due to global warming at the end of the last glaciation separated the ancient Sunda continent—for millennia an extension of mainland Asia—into present-day ISEA and Mainland Southeast Asia (MSEA). These are thought to have been concentrated in three major episodes, from 15 to 13.5 ka, 11.5 to 10 ka and 7 to 8 ka (Pelejero et al. 1999). Alternative models to "out-of-Taiwan" have argued that it may have been the rapid coastal transformation and resulting land-loss (Solheim 2006) that had the most profound effect on genetic patterns in the region, rather than a more recent expansion from Taiwan (Oppenheimer 1998; Oppenheimer and Richards 2001; Soares et al. 2008, 2011) However, although such models attempt to explain the current population structure in ISEA, they have been less successful in incorporating the linguistic evidence suggesting an Austronesian origin in Taiwan (Barker and Richards 2013).

On the genetic side, many seemingly contradictory results have been published in recent years, shifting the perspective back and forth between a strong Neolithic expansion and minor or non-existent dispersals from Taiwan. The two-layer colonization model (Pleistocene colonization and mid-Holocene "out-of-Taiwan" expansion) (Bellwood 1997) often remains the lens through which data are interpreted. Thus, genetic variation is often categorised either as autochthonous (first colonization or "Melanesian") or as a later Asian input interpreted as "Austronesian" (Friedlaender et al. 2008; Kayser et al. 2008b). The "Asian" signal on the mtDNA is generated by the so-called "Polynesian motif" (Delfin et al. 2012; Melton et al. 1995; Redd et al. 1995; Soares et al. 2011; Sykes et al. 1995; Trejaut et al. 2005), which approaches 100 % in many Remote Pacific islands.

However, Soares et al. (2011) estimated an arrival of this clade (or its ancestor) in Near Oceania ~6 to 10 ka. Although the motif has an ultimately mainland Asian ancestry (in haplogroup B4a1a) sometime in the last 10–20 ka (Soares et al. 2011), it was already well established in Near Oceania by the mid-Holocene. This implies that the contribution of a Neolithic "out-of-Taiwan" migration to Remote Pacific Islanders is negligible in the mtDNA [as well as Y-chromosome (Capelli et al. 2001; Kayser et al. 2000)] variation in the last 3 ka. But, for ISEA, too, the picture is far from consistent with an "out-of-Taiwan" demic expansion. The largest surveys consistently suggest a far more complex picture than the two-layer model (Capelli et al. 2001; Hill et al. 2007; Karafet et al. 2010; Trejaut et al. 2014; Tumonggor et al. 2013). Sea-level rises probably shaped much of the genetic structure of ISEA (Hill et al. 2007; Karafet et al. 2010), with major dispersals originating in what is now the mainland [including mtDNA

[17] Department of Archaeology and Natural History, College of Asia and the Pacific, The Australian National University, Acton, Canberra, ACT 2601, Australia

[18] MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, Scotland

[19] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rua São Francisco Xavier, Rio de Janeiro 20550-900, Brazil

[20] Faculty of Medicine, University of Porto, Al. Prof. Hernâni Monteiro, 4200-319 Porto, Portugal

[21] Institute of Human Sciences, School of Anthropology, University of Oxford, The Pauling Centre, 58a Banbury Road, Oxford OX2 6QS, UK

haplogroup B4a1a (Soares et al. 2011)] as well as across what is now ISEA [including haplogroup E (Soares et al. 2008)].

Genome-wide data have also led to challenges to the "out-of-Taiwan" model, albeit with caveats (discussed below). The Pan-Asian SNP Consortium (Abdulla et al. 2009) suggested that the diversity of Taiwanese aboriginals is likely a sub-set of the ISEA diversity, implying that dispersals between Taiwan and ISEA took place in the reverse direction. This would match the situation seen in mtDNA haplogroup E, inferred to have expanded in ISEA in the postglacial period and reached Taiwan within the last 8 ka (Soares et al. 2008).

Here we perform founder analyses with large new mtDNA datasets, both control-region and whole-genome sequences, and—for the first time—Y-chromosome data. Founder analysis estimates dispersal times and quantifies the contribution of each migration to the present-day population. We develop an explicit set of criteria by which to evaluate candidate "out-of-Taiwan" markers, and show that haplogroup M7c3, analysed here at the maximal resolution level of whole-mtDNAs, and found in aboriginal Taiwanese and the Philippines at moderate frequencies, but only low frequencies in ISEA and the western Pacific, fulfils these criteria almost perfectly. However, the other major candidates proposed for the "out-of-Taiwan" dispersal, haplogroups E and B4a1a, fail to meet any of them.

Single-locus studies of the uni-parental marker systems can today provide exquisite resolution, but they are, of course, subject to greater stochastic effects than the autosomal genome. We therefore here back up mtDNA and Y-chromosome variation with fresh analyses of the autosomal genome-wide structure of Southeast Asians. These multi-locus analyses support the view that the spread of the red-slipped pottery Neolithic and Austronesian languages in ISEA were indeed accompanied by dispersals of seafarers from Taiwan, but beyond the Philippines the primary mechanism for the spread of both was acculturation. In fact, a slightly earlier Neolithic dispersal from MSEA, involving paddle-impressed ceramics and possibly accompanied by Austroasiatic languages, had a substantially greater genetic impact on much of ISEA, especially in the south.

Methods

mtDNA founder analysis

Founder analysis (Richards et al. 2000) works by identifying founder types (the result of individual migration events from source to sink in the past) and then partitioning clusters derived from them in the sink population on the basis of their coalescence times, to estimate arrival times. This mtDNA founder analysis departed from reduced-median networks (Bandelt et al. 1995) of data from HVS-I (the first hypervariable segment) of the mtDNA control region (between nucleotide positions 16,051 and 16,400), usually augmented by haplogroup-diagnostic coding-region variants.

For our sink region, we analysed 2216 mtDNA sequences from Island Southeast Asia (556 from the Philippines, 340 from Borneo, and 1320 from the rest of Indonesia), including 320 new sequences from ISEA (183 from Sabah, Brunei and Kalimantan in Borneo) and published data (Table S1). For the source we included 6070 Chinese sequences, 1429 from MSEA, 827 aboriginal Taiwanese and 4573 sequences throughout North and Central Asia. We included additional data from Malaysia (519 Malays and 308 *Orang Asli*), 55 unpublished sequences from Singapore and published data from other regions to further resolve the phylogenetic networks (Table S2). A second database was created with the aim of performing a founder analysis for Remote Oceania. Adding to the sequences above, datasets for New Guinea (846), the Karkar Islands (47), the Solomon Islands (258), the Bismarck Archipelago (1005) and Bougainville (255) were included in the new source population. Sequences from Vanuatu (130) and throughout Polynesia (148) were included in the sink population (Table S3).

We analysed the data phylogenetically haplogroup by haplogroup, and carried out the founder analysis as before (Rito et al. 2013), including a 200-year scan as a preliminary step. We estimated errors using the approach of Saillard et al. (2000), to allow for non-star-like founder clusters. In this approach, we replaced the number of samples in the ρ estimation by an effective number of samples based on the number of samples that would be present in a completely star-like network associated with the same level of uncertainty as we have implemented before (Soares et al. 2012) including at the whole-population level for South African populations (Rito et al. 2013). We employed a mutation rate of one mutation every 16,677 years for the range 16,051–16,400 (Soares et al. 2009) in the founder analysis.

We regard the scan as a heuristic approach to detecting and dating peaks of immigration, and the partition analysis as an attempt to quantify and place confidence limits upon them. Following examination of the heuristic scan, we then used archaeological evidence to finalize the dates chosen for the partition analysis, which should compensate for any systematic bias in the HVS-I mutation rate we adopted. 4.5 ka approximates the earliest likely arrival of both the putative "out-of-Taiwan" Neolithic and the paddle-impressed-ware Neolithic from MSEA in ISEA and is likely to be conservative in the case of the former—an age of 4.5 ka is more persuasive for the influence of MSEA

than Taiwan, which more likely postdates 4 ka (Anderson 2005; Spriggs 2007). The putative episode of sea-level rise at ~8 ka (Pelejero et al. 1999) was initially the least well-established of the three episodes (Blanchon and Shaw 1995), but a rapid rise between 7.4 and 6.5 ka has recently received clear support (Bird et al. 2010). The existence of human dispersals at around this time is, however, supported by whole-mtDNA genome evidence from haplogroup E in ISEA and Taiwan (Soares et al. 2008), and by the founder analysis scan. For the primary settlement, we used 50 ka as a rough approximation; evidence from Niah Cave and New Guinea suggest an age upwards of 49 ka (Hunt et al. 2007; Summerhayes et al. 2010).

Y-chromosome founder analysis

We enhanced the resolution of previously available data (Capelli et al. 2001) by augmenting with 10 more Y-STR (short tandem repeat) markers and five more SNPs. Adding to the ten previous Y-chromosome STRs (DYS388, DYS393, DYS392, DYS19, DYS390, DYS391, DYS425, DYS426, DYS389I and DYS389II), we included an additional ten: DYS460, DYS461, DYS438, DYS448, DYS458, DYS437, DYS439, H4, A10 and DYS635. We typed the ten Y-STRs in a multiplex that also included the previously typed Y-STRs, DYS388 and DYS425 (Capelli et al. 2001), to control against mixing of samples across the studies, and as a result we excluded several samples. The DYS426 marker gave inconsistent typing results across the dataset and we therefore excluded this locus from further analyses. Table S4 shows a list of the primers used in the multiplex. We ran the samples on an ABI 16-capillary 3130XL DNA Analyser at the University of Leeds.

Five SNPs were added that included: M230, which allowed the identification of the Pacific haplogroup S; M324, which allowed the identification of O3a against O3* in the previous dataset (Capelli et al. 2001); M50, which allowed the identification of subclade O1a2 within O1a, which were to prove highly relevant to identifying "out-of-Taiwan" founders; and M38 and M208, which permit the identification of subclades C-M38 and the Pacific C-M208 in samples previously labelled C* haplogroup only (Capelli et al. 2001). We typed M230, M324 and M208 using restriction analysis (Table S5 indicate primers and enzymes used), and M50 and M38 using denaturing high-performance liquid chromatography (dHPLC), as described previously (Underhill et al. 2000). We ran the samples on an automated 3500HT Wave (Transgenomic) dHPLC instrument and analyzed the results with the Navigator software. Fig. S1 summarises the SNPs analysed and the distribution of the clades identified across the sampling locations.

We used the SNP data to test the reliability of the STR analyses by constructing a median-joining network (Bandelt et al. 1999) of the overall Y-STR data. Initially we calculated a highly reticulated network using the same standard weight (10) for all STRs. We registered the number of occurrences of each STR change, and used these values to generate a graduated weighting scheme for a new network reconstruction (weight of 1 for the STR showing the highest number of occurrences and a weight of 10 for the lowest). We repeated this process until the phylogeny became stable with further iterations. The final weighting scheme gave DYS458 a weight of 1, while DYS388, DYS425, DYS437 and DYS438 were each given a weight of 10. The new STR network showed a remarkable congruence with the clustering of SNP-defined lineages (Fig. S2), with just a few minor exceptions (most particularly, ten C-M38 lineages that separated from the main cluster), showing that, although the SNP tree is limited, the Y-STR data, suitably analysed, can provide an unbiased and reliable reconstruction of ancestry. STR markers are, of course, far more unstable than bi-allelic variants but, if we consider the difficulties associated with generating a reliable tree of mtDNA haplogroups based on HVS-I data alone, we can assume we have a comparable resolution to the maternal founder analysis counterpart—possibly higher.

Zhivotovsky et al. (2004) calculated an effective mutation rate of 6.9×10^{-4} mutations in 25 years per Y-STR, or an average Y-STR mutation rate of 2.76×10^{-5} mutations per year, which is slower than estimated Y-STR pedigree rates by an order of magnitude (Gusmão et al. 2005). Since our dataset included a star-like clade that is confined to Remote Oceania (Fig. S3), and considering that the colonization time of this region is well established by radiocarbon evidence, we opted for re-calibrating the specific 19 Y-STR average mutation rate using the settlement of the Remote Pacific as the calibration point. To this end, we constructed a network of haplogroup C-M208 and applied weights obtained previously in the general Y-STR network (Fig. S4) to the network. C-M208 shows three main branches, two present in Madang and Vanuatu and a third present only in Remote Oceania (excluding Vanuatu) from Fiji to French Polynesia. The expansion into Remote Oceania began as early as ~3.3 ka, indicated by radiocarbon estimates at a Lapita site in Vanuatu (Bedford et al. 2006), while Western Polynesia (Tonga and Samoa) was colonised ~2.9 ka (Rieth and Hunt 2008). The star-like subclade, C-M208, was not present in Vanuatu and we therefore assumed an estimate of 3000 years for the age of this subclade, whose westernmost location was Fiji [see discussion in Clark and Anderson (2009)]. We estimated an average Y-STR mutation rate of 4.08×10^{-5} mutations per year, meaning nearly 50 % faster than the one calculated by Zhivotovsky et al. (2004). We must emphasize that this average mutation rate corresponds specifically to this group of STRs, and also that we are dealing with a very recent human event in the

calibration. Our estimated mutation rate is still substantially slower when compared with father–son transmission studies (Ballantyne et al. 2010; Gusmão et al. 2005).

As for the HVS-I mtDNA data, we calculated networks for each clade. We rooted the networks by estimating the root through midpoint rooting and using an outgroup of the consensus STR length of the closest available clade to further pinpoint a hypothetical root. Again we selected founders using both *f1* and *f2* criteria (Richards et al. 2000). We used the Bayesian migration partition tool (Richards et al. 2000) in the same two ways: a scan of equally distant intervals of 200 years (Rito et al. 2013; Soares et al. 2012) and a model of migration with migration time windows based on the scan and archaeological and climatological data. This is the first direct application of founder analysis to Y-chromosome data.

The partition model included, as for mtDNA, migrations at 0.5 ka (for recent gene flow), 4.5 ka (Neolithic) and 8 ka (postglacial migrations) differing only in the time of the older migration (20 ka instead of 50 ka). We do not expect the ρ statistic (Forster et al. 1996) to be a good estimator of age for more ancient lineages in a highly mutating system such as STRs. Because of this, using a realistic time of first settlement based on archaeology of ~50 ka (Hunt et al. 2007) in the partition model would cause the older lineages dating to about 20 ka to be statistically allocated in the postglacial migration at 8 ka, to which they are closer, which we believe would be quite misleading. We therefore opted for including a migration at the time of the peak, 20 ka, even though the age of the peak probably does not correspond to the time of migration.

Validation: founder analyses for Remote Oceania

To evaluate the performance of the methodology, we also executed a founder analysis from ISEA/Near Oceania into Remote Oceania. The well-characterized time of expansion into the Remote Pacific islands provides a valuable framework for testing the clock and the founder analysis methodology that we employed. The expansion into Remote Oceania may have begun before 3 ka into Vanuatu (Bedford et al. 2006), but the major migration into Western Polynesia began only within the last 3 ka (Rieth and Hunt 2008). In our founder analysis using a 200-year scan we obtained the same pattern for the two criteria: a single peak at 3000 years (Fig. S5), fitting very well the archaeological data.

We obtained a similar pattern for the Y-chromosome analysis (Fig. S5). However, contrary to the mtDNA analysis, we should re-emphasize that this is very far from a completely independent check for the methodology, since the rate of the Y-STRs we employed was calibrated assuming 3000 years for the major founder clade entering Remote Oceania. It is nevertheless reassuring that the time of the peak was not affected by the inclusion of the other clades in the analysis. We should note also that the estimated mutation rate is to be employed in determining migratory fractions in ISEA and not the Pacific, so its use does not provide circular evidence, only the cross-checking.

Genome-wide analysis

We used 1251 samples taken from the Pan-Asian SNP Genotyping Database (Abdulla et al. 2009; Ngamphiw et al. 2011) (Table S6). We used the West African Yoruba data as an outgroup, as well as a South Asian group. The objective of the analysis is to compare the genome-wide patterns with the haploid marker results.

The initial dataset contained 54,794 SNPs. We pruned this dataset for linkage disequilibrium (LD) using PLINK (Purcell et al. 2007). One SNP in pairs with LD higher than $r^2 = 0.1$ was removed in windows of 50 SNPs shifted five SNPs each time, as used before (Pierron et al. 2014; Verdu et al. 2014). We used a total of 23,332 SNPs in further analyses. We employed ADMIXTURE (Alexander et al. 2009) to estimate population structure using a maximum likelihood approach, assuming different numbers (2–15) of ancestral populations or genetic components (K). A few populations, namely the "Negrito" groups in Malaysia, displayed a single private component for very low values of K that is not present elsewhere, most probably due to the effect of strong genetic drift. As the objective of the analysis was to display overall geographic patterns to provide a frame of comparison with the uniparental markers' phylogeography, and since these populations did not provide any relevant information in this regard, we opted for excluding them from the final analysis. We performed a cross-validation by inspecting the cross-validation error (CVE) in the analyses with different values of K (2–15). Theoretically, the one with the lowest CVE should be the most accurate. A graphic of the variation of the CVE obtained is shown in Fig. S6. Although the CVE does not vary much above $K = 5$, $K = 10$ displays the lowest value.

To visualize the distribution of specific mtDNA clades or autosomal components, we displayed the frequency distributions using the Kriging algorithm of Surfer 8. Data points used for the mtDNA and the genome-wide components are shown in Fig. S7.

Whole-mitochondrial genomes

We analysed lineages from across the range of variation in mtDNA haplogroup M7, but with a particular focus on the candidate "out-of-Taiwan" marker, M7c3c. We generated a total of 114 new M7 sequences, including 38 Taiwanese, 20 Vietnamese, 16 Indonesians, 12 Peninsular Malaysians,

7 East Malaysians, 6 Laotians, 4 Chinese, 4 Micronesians (from Nauru and Kiribati), 3 Bruneians, 2 Filipinos, 1 Burmese and 1 Thai. We also extracted 51 new M7 sequences from the raw data of the 1000 Genomes project (unavailable at the start of this study). We performed whole-mtDNA sequencing as previously described (Torroni et al. 2001) using an ABI 48-capillary 3730 DNA Analyser (Taipei) an ABI 16-capillary 3130XL DNA Analyser (Leeds) and an ABI 16-capillary 3100 DNA Analyser (Porto). Details on the new and published sequences used in the phylogenetic reconstruction of haplogroup M7 are indicated in Table S7. We deposited the new whole-mtDNA sequences in GenBank (accession numbers JX987440-JX987470 and KU131308-KU131390). For comparison with mtDNA haplogroup M7, we also re-analysed haplogroups B4a1 and E using all the available published whole-mtDNA genome sequences. We list the samples we used in the analyses in Tables S8 and S9.

We reconstructed phylogenies of haplogroups M7, B4a1a and E using Network 4.6 software with the reduced-median algorithm (Bandelt et al. 1995), resolving reticulations on the basis of the relative rates of the mutations involved (Soares et al. 2009). We estimated ages for the different phylogenies using both the ρ statistic (Forster et al. 1996) and maximum likelihood (ML), using the mtDNA clock of Soares and collaborators that corrects for purifying selection from the long-term phylogenetic rate of one mutation every 3624 years (Soares et al. 2009). We estimated branch lengths in ML using PAML 3.13 (Yang 1997) assuming the HKY85 substitution model with gamma-distributed rates (32 categories). We also employed a synonymous mutation rate of one substitution every 7884 years (Soares et al. 2009) using the ρ statistic (Forster et al. 1996). We have discussed in some detail previously the potential impact of mutation rate uncertainty on phylogeographic conclusions (Mellars et al. 2013) and we re-calculated the confidence intervals in similar fashion here.

We obtained Bayesian skyline plots (BSPs) (Drummond et al. 2005) using BEAST 1.4.6. (Drummond and Rambaut 2007) to detect signatures of population increment associated with the haplogroups under analysis. We employed a relaxed molecular clock (lognormal in distribution across branches and uncorrelated between them), a mutation rate of 2.514×10^{-8} mutations per site per year for the whole-mtDNA genome (Pereira et al. 2010) and the HKY model of nucleotide substitutions with gamma-distributed rates, assuming a generation time of 25 years. We compared different signatures of population growth in ISEA and aboriginal Taiwanese data for the three haplogroups analysed, M7c3c, B4a1a and E.

Given a recent age estimate for haplogroup E based on a single ancient DNA sequence (Ko et al. 2014) that diverges significantly from previous estimates for this haplogroup, we performed a new estimate that was also based on ancient DNA sequences. Ancient DNA calibration is becoming a widely used approach (Ho et al. 2007), but estimates based on a single sequence are highly divergent and present large confidence intervals (Fu et al. 2013b). Calibrating a clock with only a single ancient DNA sequence can be misleading, particularly for recent samples for which even the stochastic presence of one more or one less mutation than the expected clade average can lead to strong departures from a realistic mutation rate. For calibration purposes, we added two more ancient East Asian sequences to the haplogroup E sequence described by Ko et al. (2014). One dates to the same time-frame as the E sequence: Boshan 11, from north-east China, at 8.18 ka (Fu et al. 2013b). The second one is older: Tianyuan 1301, also from north-east China, at 39.5 ka (Fu et al. 2013a); it is important to include this in the analysis, since haplogroup E as a whole will necessarily be older than the ancient haplogroup E sequence. In the analysis, we used a tree that represents a snapshot of human diversity (Table S10), which allows a better estimate of the evolutionary parameters in mtDNA, with all the main haplogroup E branches represented. We employed a relaxed molecular clock (lognormal in distribution across branches and uncorrelated between them) using BEAST 1.4.6. (Drummond and Rambaut 2007), with a constant population size and the HKY model of nucleotide substitutions, with gamma-distributed rates.

Results

mtDNA control-region and Y-chromosome founder analyses

To investigate the genetic input into ISEA through time, we carried out founder analyses with both mitochondrial DNA (mtDNA) control-region sequences—for the maternal line of descent—and Y-chromosome variation using a 19 Y-STR dataset within SNP-defined lineages—for the paternal line of descent. Founder analysis is a quantitative phylogeographic approach developed to evaluate the diversity of lineage clusters that has arisen within a particular geographic sink region (in this case, ISEA), following migration from a specified (assumed) source region (in this case, MSEA/China/Taiwan). Using the molecular clock to convert to time depth, these values are a proxy for the minimum arrival age of each founder cluster in the sink (Richards et al. 2000).

For maternal lineages, the 200-year scan of founder lineages dispersing into ISEA (Fig. 1a) identified two major coalescence peaks (corresponding to bursts of immigration) under the two criteria we employed, *f1* and *f2* (Fig. 1a) (Table S11), at 4.6–4.8 ka and at 8–10 ka, respectively. We

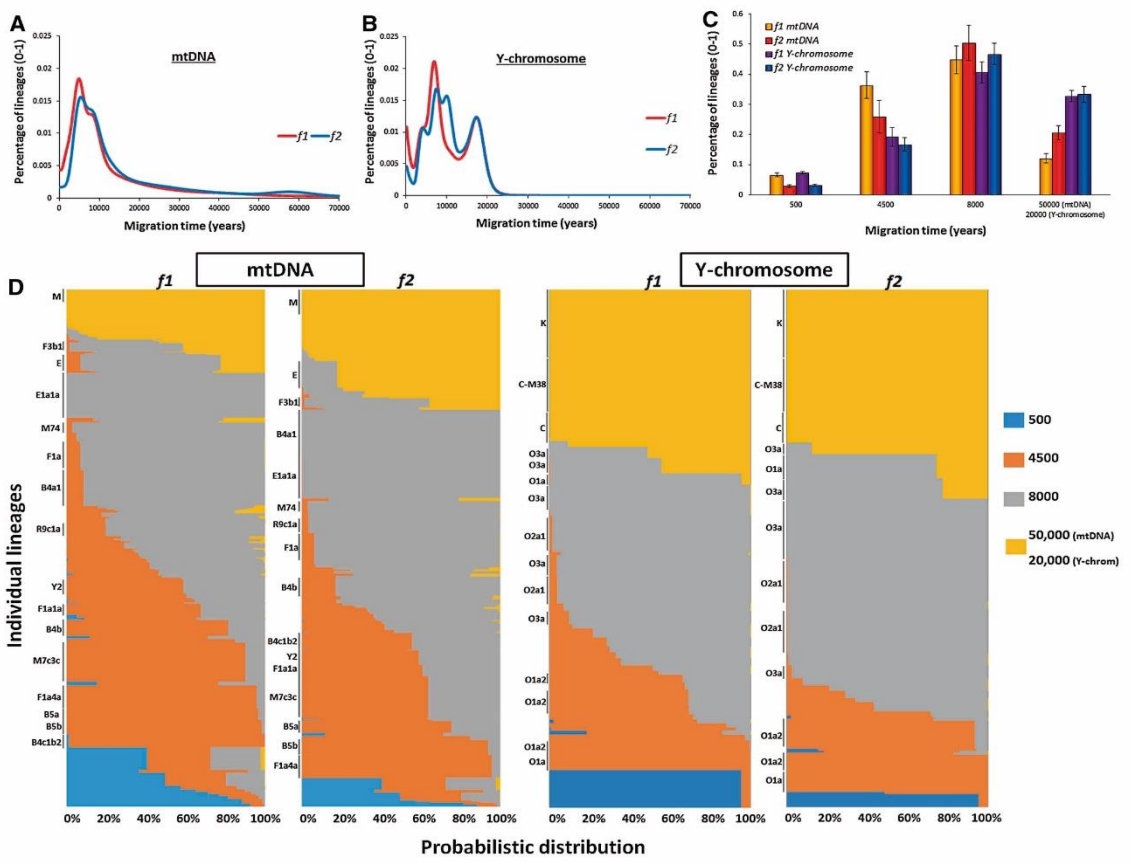


Fig. 1 Founder analysis results for ISEA, assuming Taiwan as source, for mtDNA (female lineages) and Y-chromosome variation (male lineages). **a** Probabilistic distribution of mtDNA founder clusters across migration times scanned at 200-year intervals from 0 to 70 ka, using two criteria for founder identification, *f1* and *f2*; **b** probabilistic distribution of Y-chromosome founder clusters across migration times scanned at 200-year intervals from 0 to 70 ka, using two criteria for founder identification, *f1* and *f2*; **c** proportion of founder lineages in a four-migration model for mtDNA and Y-chromosome variation using two criteria for founder identification, *f1* and *f2*; **d** probabilistic distribution of each individual lineage in mtDNA and Y-chromosome variation in a four-migration model chromosome using two criteria for founder identification, *f1* and *f2*. Individual founder clusters with more than 2 % frequency in overall ISEA (sink populations) are indicated at the left-hand side of each *plot*

also observed a slight hump ~55 ka with the *f2* criterion alone.

For Y-chromosome variation (Fig. 1b), we obtained very similar peaks with both criteria (one at 4–5 ka and a second at ~8 ka). Rather remarkably, the two main peaks in two different genetic systems with distinct mutation rates and estimated using two distinct founder criteria are consistent across each of these different analyses. In addition, we observed an increment representing very recent migrations and, with the *f2* criterion, a further extra peak at 10–11 ka. This peak might signal the second well-defined episodic flood immediately after the Younger Dryas (Pelejero et al. 1999). We did not include it in the migration model, however, for two reasons: it was detected only under a single criterion and with one genetic system; and, in any case, founders at this peak will be included statistically in the ~8 ka migration that overall can be defined as postglacial migrations. The oldest arrivals here date to ~20 ka, largely haplogroup K and C lineages. This may well correspond with the ancient minor peak for mtDNA; we expect ρ dating with STRs to provide severe underestimates for ancient clades because of mutational saturation. However, for the present analysis this is a minor issue, since we are concerned primarily with events in the Holocene. Particularly in the case of K, an older age than this could be expected, considering that K probably evolved in the region since the

first settlement as displayed by the high prevalence of K* and K subclades in the ancient Sahul populations, including Aboriginal Australians (Hudjashov et al. 2007).

We then partitioned the founders in ISEA using a migration model informed not only by the scan results in the two genetic systems, but also archaeological and palaeoclimatological evidence, to quantify the contribution of each immigration event to the extant mtDNA and Y-chromosome gene pools in ISEA. The model from mtDNA data here assumes migrations at 4.5, 8 and 50 ka, corresponding to Neolithic immigration, postglacial expansions and first Pleistocene settlement. We assumed a further dispersal at 0.5 ka to allow for any recent/historical gene flow.

For Y-chromosome variation, we used a more recent age of 20 ka to cover the more ancient migrations, as mentioned above. However, the matching of peaks at 4–5 ka and 8–10 ka for both the paternal and maternal line of descent is striking. The overall contribution at each proposed migration time for each of the two founder criteria in the mtDNA and Y-chromosome variation is shown graphically in Fig. 1c, d. The mtDNAs coalescing at the time of the first settlement (~50 ka) accounted for ~10 to 20 % of modern mtDNA lineages in ISEA. Note that many lineages from the ancient Sunda continent would very likely be present across both ISEA and MSEA, which were only finally separated by sea-level rise ~8 ka. However, MSEA is a source region in this analysis, so this value in the founder analysis corresponds to ancient lineages private to ISEA only. In the mtDNA analysis, lineages descending directly from the haplogroups carried by the first settlers correspond to M*, N*, R* and possibly haplogroup F3 (Fig. 1d). Although a recently published ancient mtDNA haplogroup E sequence (Ko et al. 2014) was used to suggest a Taiwanese source for this clade, an early origin in ISEA (Soares et al. 2008) remains more likely, as discussed below. At this ancient time-frame, Y-chromosome lineages (with STR ρ dating) are uninformative due to saturation, but haplogroups K* and even C may date to the first colonization at that time. These are above 30 % in the Y-chromosome analysis.

Overall, the migration at ~8 ka contributes the most lineages to the current gene pool of ISEA with a fraction of ~40–50 % in both mtDNA and Y-chromosome variation (Fig. 1c). We stress again that, statistically, this migration time could include lineages entering ISEA throughout the period of sea-level rises, from 14 to 8 ka, covering all three flooding episodes (Pelejero et al. 1999). This partition probabilistically includes major and well-studied haplogroups such as B4a1a (Soares et al. 2011), subclades of haplogroup E (Soares et al. 2008), F1a*, and subclades of haplogroup M shared between ISEA and MSEA, with B4a1a and E the major contributors. In Y-chromosome variation, this migration includes most clusters within haplogroups O2a1 and O3 and a subclade of O1a (Fig. S4), matching to some extent the results of Karafet et al. (2010) indicating that O1a* entered ISEA before the Neolithic. We should note that in our recent Y-chromosome survey (Trejaut et al. 2014), O2 and O3 clades declined in frequency moving north from ISEA towards Taiwan, the opposite of what one might expect from an "out-of-Taiwan" movement. A previous survey (Karafet et al. 2010) also suggested that O3, O2a1 and O1a* entered ISEA from the mainland before the Neolithic period.

The contribution at the time of the Neolithic, at 4–5 ka, varied with the criterion and the genetic system, but 25–35 % is probably the best estimate (Fig. 1c). (The *f1* criterion in mtDNA probably overestimates recent migration due to the large size of the source sample used.) Only one major founder presented significant differences between the analyses: B4b appears Neolithic in *f1* criterion and part of the postglacial migration in the *f2* criterion (Fig. 1d). This haplogroup deserves further attention in the future. The widely held model for the spread of the Neolithic in ISEA implicates expanding pre-Austronesian/Austronesian speakers from South China/Taiwan (Bellwood 1997); but in fact not all of the Neolithic founders we identify support this hypothetical "out-of-Taiwan" dispersal. A large fraction of Neolithic mtDNA founder clusters from haplogroups B5a1 and F1a1a (~10 % out of the 25–35 % Neolithic lineages in the analysis) appear to have originated in MSEA, and are rare or absent in either Taiwan or the Philippines.

Our results therefore suggest that mid-Holocene Neolithic immigration into ISEA was in part via MSEA, temporally associated with spread of basket-marked and carved paddle-impressed pottery, which appeared across MSEA as early as red-slipped pottery appeared in Taiwan (Bulbeck 2011), and possibly involving speakers of Austroasiatic languages (i.e. Anderson's "Neolithic I") (Anderson 2005). The mtDNA haplogroups M7c3c, Y2, F1a4a, B4c1c and possibly B4b (which shows contrasting patterns under the two criteria) may, however, represent genuine "out-of-Taiwan" clades in ISEA. These founders are all derived from Chinese-mainland source haplogroups, and within Austronesian-speaking populations they have a higher overall frequency in Taiwan and the Philippines (Fig. 2a). This input, at ~20 %, lends support to a modified, small-scale "out-of-Taiwan" model [Anderson's "Neolithic II" (Anderson 2005; Donohue and Denham 2010)], proposed to explain the appearance of red-slipped pottery in relation to the early dispersal of Austronesian languages.

On the male line of descent, the Neolithic contribution is lower (15–20 %) but, since MSEA is not represented in the Y-chromosome dataset, all these Neolithic founders are likely to represent the putative "out-of-Taiwan" dispersal, mirroring closely the ~20 % "out-of-Taiwan" founders for

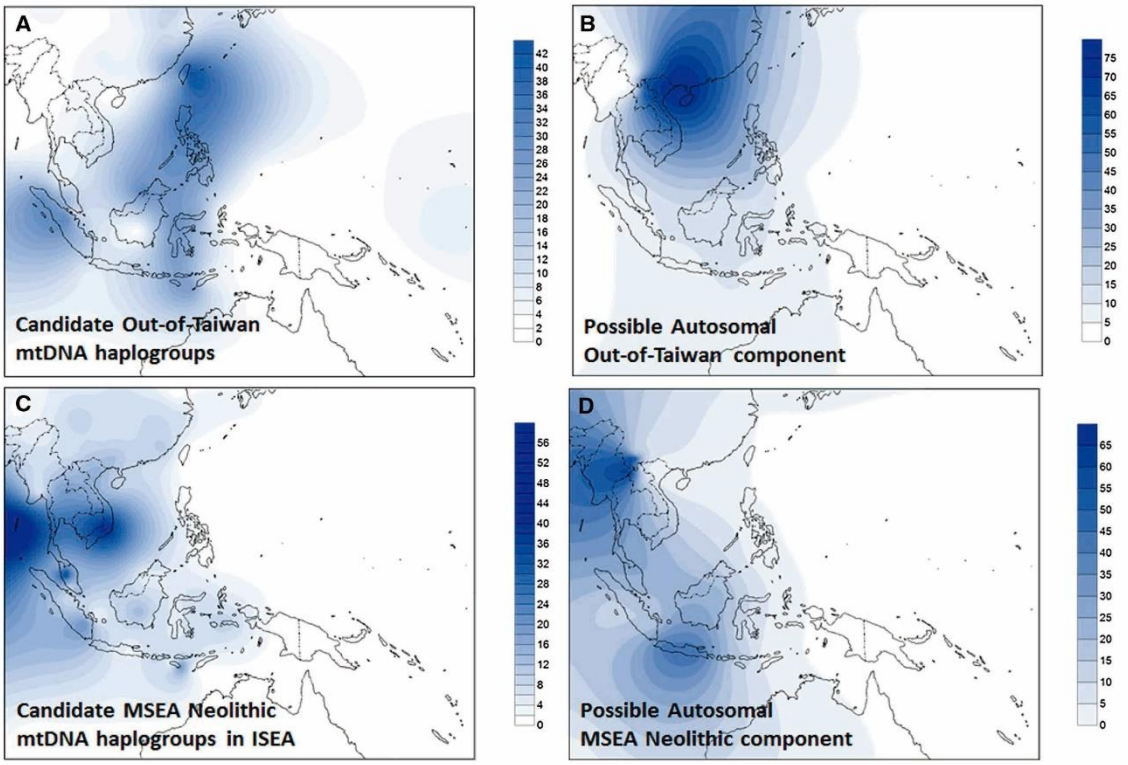


Fig. 2 Frequency map of probable Neolithic markers (lineages argued to track one or other of the dispersals associated with Neolithic ceramics) in mtDNA and genome-wide data. **a** Pooled frequency of candidate "out-of-Taiwan", "Neolithic II" mtDNA haplogroups, based on founder analysis. **b** Possible "out-of-Taiwan", "Neolithic II" component in the genome-wide data when considering 10 ancestral populations in the ADMIXTURE analysis. **c** Pooled frequency of candidate MSEA "Neolithic I" haplogroups in ISEA. **d** Possible MSEA "Neolithic I" component in the genome-wide data when considering 10 ancestral populations in the ADMIXTURE analysis. The outline map was obtained from http://www.outline-world-map.com

mtDNA. Most of O1a and all of O1a2 likely represent signals of Neolithic migrants from Taiwan, confirming earlier suggestions (Karafet et al. 2010; Trejaut et al. 2014). A portion of O3a (~10 % in the *f1* criterion) was also partitioned into the Neolithic in our analysis.

Corroboration of founder analyses with genome-wide evidence

We next compared these results with patterns observed in autosomes, using genome-wide data from the Pan-Asian SNP Genotyping Database (Abdulla et al. 2009; Ngamphiw et al. 2011) and the ADMIXTURE software.

At a more basal level, the first that seems anthropologically and genetically potentially valid ($K = 5$, which includes African, South Asian and Near Oceanian components in purple, white and blue) (Fig. 3a), the East Asian autosomal data separate into a Southeast Asian component (green) with a focus on the ancient Sunda continental shelf (MSEA, Sumatra, Java and Borneo) that varies from ~80 % around Borneo and drops in frequency as one moves north, and a Chinese/Northeast Asian component (red), which varies between 100 and 60 % in mainland China. Frequencies of the latter in Taiwan (~30 %) and Southeast Asia (5–30 %) match the mtDNA picture of Neolithic-age Chinese gene flow into ISEA (Fig. S8; cf. Fig. 2a). It is, however, difficult to directly connect a given component in ancestry analysis with a given demographic occurrence. One could calculate the time of admixture, but admixture ages are not necessarily indicative of time of migration (Lipson et al. 2014). In addition, the ages calculated are sometimes dubious and under-estimated as the estimated time of split between Europeans and New Guineans suggests (Wollstein et al. 2010).

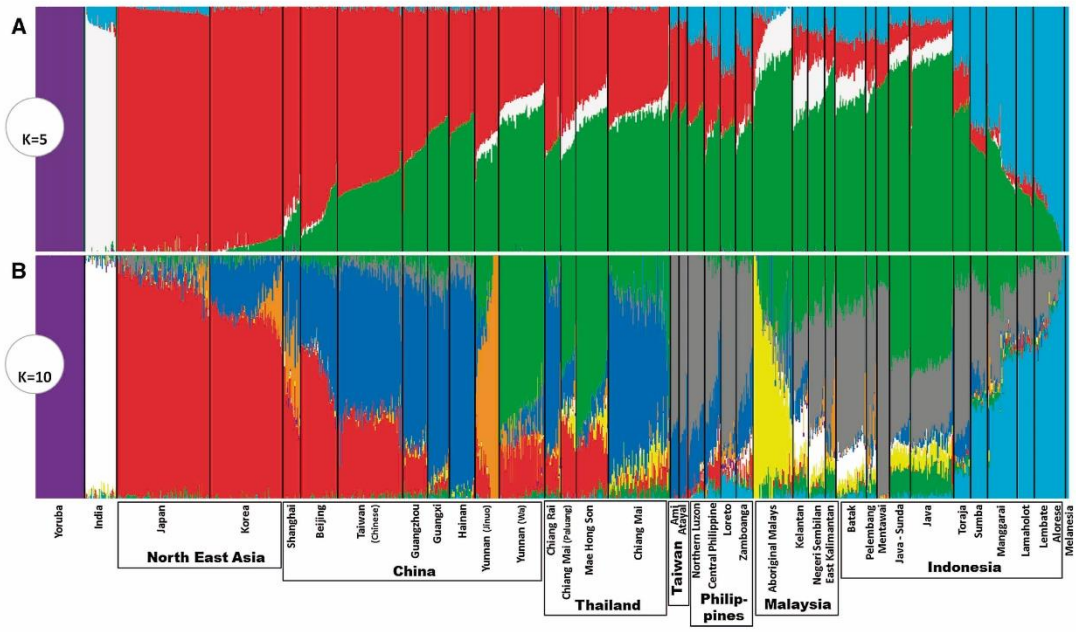


Fig. 3 Reconstruction of ancestry in Asian populations using ADMIXTURE. Considering **a** five ancestral populations ($K = 5$) and **b** 10 ancestral populations ($K = 10$)

Analyses from $K = 6$ to $K = 15$ generate additional components by further sub-dividing these Northeast and Southeast Asian components, whilst maintaining the African, South Asian and Melanesian/Near Oceanian components intact across the analyses. The autosomal estimate with ten ancestral populations, theoretically the best estimate of ancestry for the data as it has the lowest cross-validation error, includes seven components with discernible frequencies in at least one location in Austronesian-speaking populations (Fig. S9). The South Asian component (white) is present at low frequencies only in the Malay Peninsula and Sumatra, matching the historical record (Manguin et al. 2011). The Northeast Asian component (red) is seen at appreciable frequency only in the Philippines (at only very low rates). A Near Oceanian component (pale blue) dominates many of the populations of Eastern Indonesia, as expected. Two minor components (dark green and yellow) are virtually specific to ISEA, mainly in what was western Sundaland (Java/Sumatra/Malay Peninsula), with one (yellow) markedly elevated in Aboriginal Malays.

One important component (grey) is both specific for Austronesian-speaking populations and highly frequent across ISEA (Fig. S9). It reaches 60–70 % in the two aboriginal Taiwanese groups in the sample—the equivalent cluster in Pan-Asian SNP data approaches 100 % (Abdulla et al. 2009)—peaking in our dataset in the Philippines, Sumatra and Sulawesi (70–90 %), and is virtually absent from Continental Asia, suggesting an insular origin. Comparison between the analyses with five and ten ancestral populations also suggests that this was part of the larger Southeast Asian component in the former. Considering the major postglacial signal observed in mtDNA and Y-chromosome variation in both our founder analysis and in earlier analyses (Hill et al. 2007; Karafet et al. 2010; Soares et al. 2008, 2011; Trejaut et al. 2014), and the sharing of many lineages between ISEA and Taiwan (Soares et al. 2008, 2011), this autosomal component may correspond to an ancestral cluster common to both Taiwan and ISEA that was established before the hypothetical dispersal of Austronesian. Even if we consider that there is likely a signal of Austronesian expansion "out-of-Taiwan" in the genome-wide data (see below), this component, which is most frequent in Taiwan, the Philippines, the Mentawai Islands and Sulawesi, disparate islands at opposite extremes of the Sunda shelf, could explain why a maximum likelihood population tree of the Pan-Asian SNP data indicated Taiwan as an offshoot of ISEA diversity (Abdulla et al. 2009). Such population trees only depict broad patterns and, although a minor component could show an ancestry in Taiwan when compared with ISEA, the most frequent component could show the overall opposite ancestry.

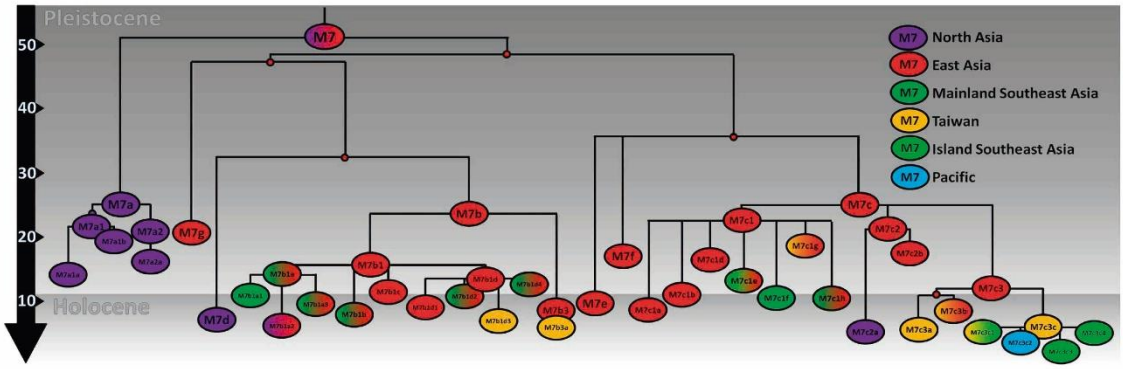


Fig. 4 Schematic tree of haplogroup M7. The tree is scaled using maximum likelihood and a time-dependent molecular clock for whole-mtDNA genomes

Two autosomal components that might signal Neolithic dispersals can be compared with the patterns obtained from Neolithic founder candidates in the mtDNA analysis. One of these components (paler green in Fig. 3b) is frequent in MSEA/Southwest China (up to ~70 %) and varies from 5 to 40 % in Indonesia (Fig. 2d), but is absent from Taiwan and rare in the Philippines. It is probably a relatively recent arrival as it is not evenly distributed across ISEA. The MSEA Neolithic candidates in the mtDNA also show a strong peak of frequency in MSEA and frequencies of 5–30 % in Indonesia, but are rare in the Philippines and Taiwan (Fig. 2c). We can also match these distributions with the presence of basket-marked and carved paddle-impressed pottery: in Sarawak, assemblages at ~4.5 ka with carved cord-wrapped or basketry-wrapped paddle-impressed pottery (Bellwood 1997; Bulbeck 2008) show the influence of an early Neolithic from MSEA in Western Indonesia.

The final component (dark blue in Fig. 3b) has a high frequency in South China (Fig. 2b) and is also seen in Taiwan at ~25–30 %, in the Philippines at ~20–30 % (except in one location which is almost zero) and across Indonesia/Malaysia at 1–10 %, declining overall from Taiwan within Austronesian-speaking populations. The mtDNA candidates for "out-of-Taiwan" markers (Fig. 2a) also show an overall frequency of up to ~35 % in Taiwan and the Philippines, but are almost absent in parts of Borneo, Java and Eastern Indonesia. Sumatra superficially presents a more discordant picture between genome-wide and mtDNA results, but the sampling of the Pan-Asian SNP dataset involves only Batak people whilst our mtDNA sampling involved the wider Sumatran population. We should also bear in mind that the genome-wide sampling lacks major areas of ISEA, including the whole of Borneo.

Therefore, the overall picture from the ADMIXTURE analysis with 10 ancestral populations where the cross-validation error was the lowest, is concordant with the mtDNA and Y-chromosome pattern, with a minor Neolithic input from MSEA, probably immediately preceding a Neolithic input from Taiwan (Anderson 2005) that had a strong demographic impact in the Philippines, but a much more minor genetic input elsewhere in the Indo-Malaysian Archipelago.

Confirmation with whole-mtDNA genome data

Although providing much larger sample sizes, the low phylogenetic resolution of mtDNA HVS-I data can create problems for phylogeographic analyses such as founder analysis, for example by conflating distinct founders. In parallel, we therefore checked the phylogeographic signal with the much better resolved whole-mtDNA genomes for the major "out-of-Taiwan" haplogroup in the founder analysis, M7c3c. In particular, we wished to compare the results for M7c3c with the two putative postglacial signals for haplogroups E and B4a1a (Soares et al. 2008, 2011).

Haplogroup M7 dates to just over 50 ka. An overall mainland Eastern Asian distribution is clear for the M7 phylogeny (Fig. 4; full tree in Supplementary Material 2). There are two basal branches, M7a, which displays a strong Northeast Asian ancestry centred on Japan and Korea, and a second major clade encompassing M7b, M7c, M7d, M7e, M7f and M7g, which we refer to as M7b′c′d′e′f′g. This splits into two further major subclades, M7b′d′g and M7c′e′f both with an East Asian ancestry.

The overall phylogenetic and phylogeographic pattern is strikingly clear: both aboriginal Taiwanese and Island Southeast Asian-specific lineages are close to the tips of an overall mainland Eastern Asian distribution. The major subclade of M7b3, M7b3a, is only present in Taiwan and ISEA. It is frequent in Taiwan (at ~10 %) and considering

its age (~6 ka) seems likely to have arrived in Taiwan with the rice Neolithic from South China; but it is vanishingly infrequent across ISEA. In M7b1, M7b1d3 is also restricted to Taiwan, and with a similar age may also have arrived from China with the Neolithic, but again it is virtually absent from ISEA.

In M7c′e′f, the three subclades branch from a single node and all show evidence of East Asian ancestry. Within M7c, M7c3 is by far the most frequent and the only one to disperse significantly into Taiwan and ISEA. This clade probably had an origin in South China, with several subclades also present in Taiwan. Its major subclade, M7c3c [M7c1c in Hill et al. (2007)], here re-dated with whole-mtDNA genomes to ~5 ka, is restricted to Austronesian-speaking populations (both Taiwan and ISEA). Given the presence of other subclades of M7c3 in Taiwan and South China, the most probable source for M7c3c is in Taiwan (amongst M7c3 arrivals from China, again perhaps with the rice Neolithic), with subsequent dispersal into ISEA. Several subclades of M7c3c exist throughout Taiwan and ISEA, and there is also one in the Pacific (M7c3c2, found in both Micronesia and the Solomon Islands), dating to less than 3 ka. This pattern confirms M7c3c as a strong candidate for an "out-of-Taiwan" marker, as indicated by the HVS-I founder analysis.

We can contrast this distinctive pattern with the distribution of haplogroups B4a1a and E, both of which are—like M7c3c—largely restricted to insular, Austronesian-speaking populations. For that reason they have been proposed as candidates for "out-of-Taiwan" markers, but neither shows a direct ancestry in South China. We propose here a set of phylogeographic parameters that we expect to see fulfilled in a clear-cut "out-of-Taiwan" marker:

(a) If the haplogroup was carried into Taiwan from South China by rice-agriculturists ~6 to 8 ka, the dispersal's timing should be bracketed by the age of the ancestral clade seen in South China (upper bound) and the insular Austronesian-specific subclade (lower bound);
(b) the insular and Austronesian-specific subclade should date to after the arrival of rice-agriculturists from China ~5.5 ka, but before the "out-of-Taiwan" migration ~4.5 ka;
(c) the founder age in ISEA for the subclade should date to ~4.5 ka, the time of the "out-of-Taiwan" dispersal;
(d) the founder age from Taiwan/Philippines to the rest of ISEA should be lower than the date of the "out-of-Taiwan" migration, ~4 ka; and
(e) the expansion of the clade in Taiwan should predate the expansion in ISEA.

We evaluated each of these points in turn (Fig. 5; Table S12; note that taking into account mutation-rate uncertainty, as documented in Table S12 does not alter the conclusions). First, we consider the ML ages of key subclades, then founder ages, and finally Bayesian skyline plot (BSP) expansion time estimates. Regarding (a), B4a1a appears in Austronesian-speaking populations between 14.7 [11.0; 18.5] ka, the age of the continental ancestral clade B4a1, and 9.9 [5.5; 14.5] ka, the age of B4a1a; haplogroup E appears between 39.2 [26.9; 52.0] ka, the age of ancestral M9, and 24.0 [14.5; 33.9] ka; and M7c3c appears between 11.8 [3.9; 20.2] ka- the age of M7c3- and 5.2 [4.0; 6.5] ka. Only M7c3c clearly fits an arrival in Taiwan in line with the "out-of-Taiwan" model. B4a1a cannot be completely ruled out from these estimates, given the 95 % confidence interval of the age estimate, but it is nevertheless very unlikely (Fig. 5a, b).

Point (b) stipulates that the insular subclade should originate after the hypothetical arrival of rice-agriculturists in Taiwan and before the dispersal "out-of-Taiwan". M7c3c, at 5.2 [4.0; 6.5] ka, follows this pattern; B4a1a, at 9.9 [5.5; 14.5] ka, and haplogroup E, at 24.0 [14.5; 33.9] ka, both suggest an earlier origin within currently Austronesian-speaking populations.

Taking point (c), an average founder age for M7c3c from Taiwan into ISEA is 4.4 [3.2; 5.7] ka, matching the 4.5 ka prediction of the "out-of-Taiwan" model. Haplogroups E and B4a1a yield 8.8 [6.0; 11.6] ka and 7.3 [5.2; 9.4] ka, respectively, suggesting earlier postglacial expansions. When including the Philippines along with Taiwan as part of the source for the dispersal—point (d)—the founder for haplogroup M7c3c dated a little lower at 4.2 [2.5; 5.9] ka—a striking match to the hypothetical Austronesian arrival in the Indo-Malaysian archipelago. Haplogroup E, by contrast, yielded 6.4 [4.8; 8.0], and the B4a1a point estimate actually increased to 8.5 [4.8; 12.3] ka, when compared with the previous founder age estimate into ISEA as a whole, clearly indicating that the "out-of-Taiwan" assumption of the founder model in this case is likely to be false.

Finally (e), we used BSPs to estimate the expansion time of each haplogroup. Figure 5c–e show the increment or rate of expansion (corresponding skyline plots in Fig. S10; data in Table S13). The B4a1a data for Taiwan and ISEA (Fig. 5c) suggest a very similar time of expansion, starting ~10 ka (with a second expansion restricted to Taiwan ~2000 years ago). However, haplogroup E expanded in ISEA before Taiwan (Fig. 5d), starting ~8 ka for ISEA and ~7 ka for Taiwan. Finally, for M7 we see a first expansion in Taiwan starting ~7.5 ka, peaking at 5.2 ka, while for ISEA the expansion starts later at 5.2 ka with peak at ~4 ka, corresponding closely to the "out-of-Taiwan" model.

Therefore, haplogroup M7c3c meets all the criteria expected for an "out-of-Taiwan" marker, whereas haplogroups E and B4a1a meet none of them. Yet a haplogroup E

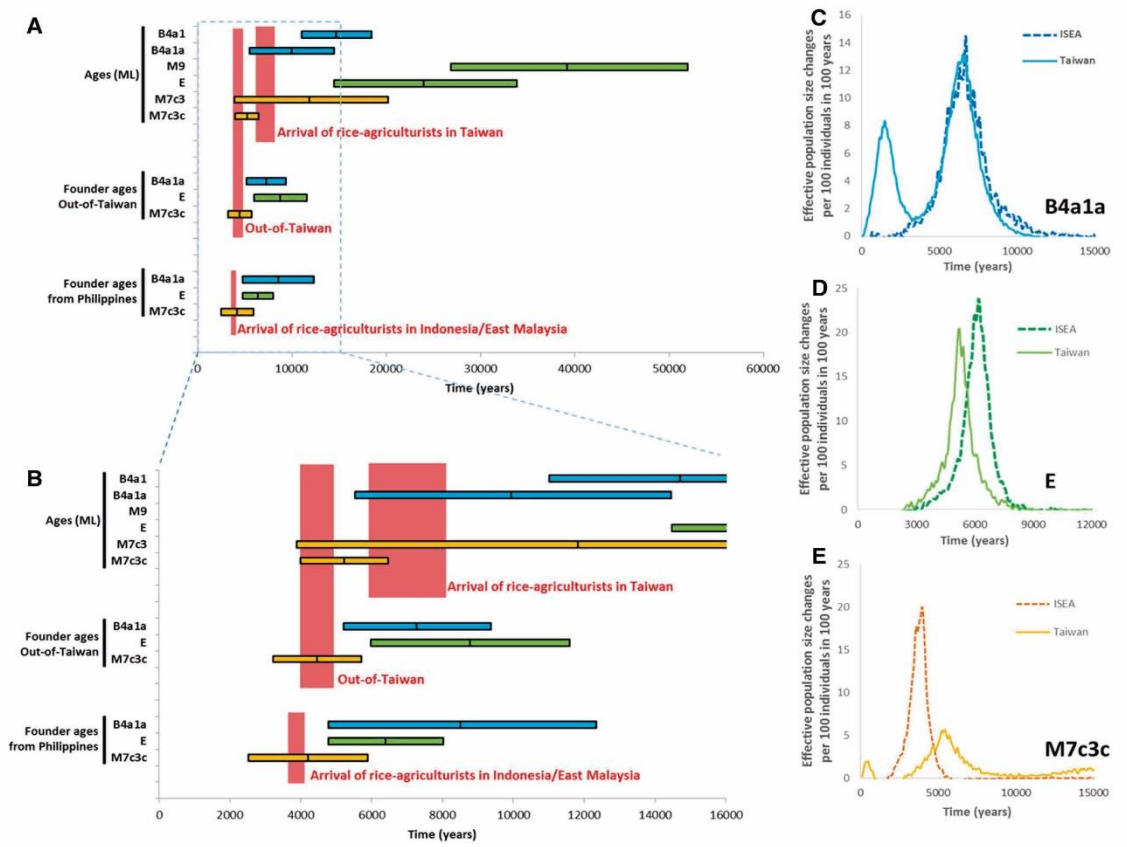


Fig. 5 Phylogeographic patterns in haplogroups M7c3c, E and B4a1a1. **a** ML ages of key clades in the test for an "out-of-Taiwan" pattern; ρ founder ages from Taiwan into ISEA; ρ founder ages from Taiwan and the Philippines into the rest of ISEA. **b** Detailed view of the most relevant time-frame for the data in **a**. **c**–**e** Increments in expansion of haplogroups B4a1a (**c**), E (**d**) and M7c3c (**e**), measured from Bayesian skyline plots as effective population size change per 100 individuals per 100 years, in Taiwan and ISEA

lineage recently recovered from human remains in the Strait of Taiwan, dating to ~8 ka, evidently represents a sequence ancestral to the E1 subclade, leading Ko et al. (2014) to suggest an origin of haplogroup E ~10 ka ago in China or Taiwan and a Neolithic migration into ISEA (based on a Bayesian analysis). This compares with our estimate for the age of haplogroup E with the time-dependent clock (Soares et al. 2009) of ~24 ka (Fig. 5). Previous age estimates based on the time-dependent clock and Bayesian ancient DNA calibrations do not differ to this extent (Fu et al. 2013b), despite some claims to the contrary. The authors of one recent estimate based on several ancient sequences claim that their estimated rate is 45 % faster than the one we estimated (Brotherton et al. 2013), but this arises from their comparing their estimated rate with our inter-specific phylogenetic rate rather than the time-dependent rate. For the time-frame of the European Neolithic and Bronze Age with which they were concerned, our curve indicates a mutation rate of 2.307×10^{-8} substitutions per site per year for the time of 6.15 ka (their oldest sample), only 4 % slower than the one they estimated. The difference would be even less for the age of their other, younger samples.

Here, indeed, we estimate an age for haplogroup E of 29.7 [18.5; 43.9] ka and an average mutation rate of 2.041×10^{-8} [1.54×10^{-8}; 2.48×10^{-8}] substitutions per site per year using a Bayesian estimate with two additional East Asian ancient DNA sequences. Given that the root of haplogroup E is seven mutations from the root of the "out-of-Africa" haplogroup M (Macaulay et al. 2005; Mellars et al. 2013) which has an average branch length to the present-day (~50,000 years) of ~20 mutations, age estimates for E more recent than ~20 ka seem implausible.

Involving haplogroup E in a wide-scale Neolithic dispersal across and out of mainland China also ignores the evidence that haplogroup E is restricted to the off-shore islands and has never been seen in any extant Chinese populations. Its age of >20 ka and insular distribution rather suggest an origin on the eastern side of the Sunda shelf. Although the early Holocene haplogroup E sequence creates a deeper link within E1, extant diversity haplogroup E diversity nevertheless remains deeper in ISEA, for both E1 and E2 (Soares et al. 2008). Moreover, a large mtDNA survey of aboriginal Taiwanese groups, which probably diverged early in Austronesian history, but were subsequently isolated and experienced drift very differently from other Austronesian populations, failed to detect any novel haplogroup E diversity, finding the same sub-set of ISEA diversity (Ko et al. 2014). The 8-ka age of the sample would place it in a period of intense postglacial expansions, due to huge sea-level changes resulting from global warming, and might be better explained as an offshoot from the south, where many lineages were lost in the postglacial period. We would caution against drawing strong conclusions from a single sample. Nevertheless, regardless of its point of origin, our analyses show that haplogroup E most probably expanded in ISEA well before the Neolithic period.

Discussion

Settlement models of ISEA that emphasize climate change and drastic shifts in the population in the early postglacial period have tended to side-step the linguistic evidence for a Taiwanese origin of the Austronesian languages (Donohue and Denham 2010). Although languages may sometimes be transmitted solely horizontally, for example by trade, this seems unlikely to explain the pattern of the Austronesian languages in ISEA as a whole, in the context of such a wide and ecologically complex region (Blench 2012). We address this issue here from the standpoint of genetic variation across the genome.

Previous results have shown a strong common ancestry between ISEA and Taiwan populations predating the pottery Neolithic period for mtDNA and Y-chromosome variation (Capelli et al. 2001; Hill et al. 2007; Karafet et al. 2010; Trejaut et al. 2014; Tumonggor et al. 2013), as well as indications that some minor lineages entered ISEA during the Neolithic. Here we show that two Neolithic waves entered ISEA, as previously suggested on the basis of pottery comparisons (Anderson 2005) and recently from autosomal analyses (Lipson et al. 2014), but that both were small-scale affairs.

The first Neolithic migration, from MSEA ["Neolithic I" in the scheme of Anderson (2005)], reflected in the distribution of haplogroups B5a1 and F1a1a and the "pale green" genome-wide component, took place ~4.5 ka and affected mainly Western Indonesia/Borneo—although it extended as far as Eastern Indonesia, particularly in the south, even reaching regions of contact with Papuan populations. A signal for this dispersal was also recently proposed by Lipson et al. (2014), although they favoured admixture with Austronesian agriculturists dispersing around the coasts of MSEA as an explanation, which our results render unlikely.

The second Neolithic wave ["out-of-Taiwan" or Anderson's "Neolithic II" (Anderson 2005)] is marked by the appearance of red-slipped pottery ~4 ka (Spriggs 2007, 2011) and impacted strongly on the Philippines (accounting for 30–40 % of current genetic diversity), where domesticated rice does indeed appear relatively early in the archaeological record (Paz 2002). However, for the rest of ISEA (the Indo-Malaysian archipelago), the demographic impact was much lower—often negligible. The overall fractions of "out-of-Taiwan" immigrants in the founder analysis for both mtDNA and Y-chromosome variation are very similar at ~15 to 20 %, suggesting that previous models inferring highly divergent male and female contributions are incorrect (similarly to the Pacific). The mtDNA haplogroup M7c3c, in particular, closely matches the expected pattern for an "out-of-Taiwan" marker.

Thus, although the Neolithic dispersal from Taiwan suggested by red-slipped pottery proves not to have been a large-scale demographic event (at least, beyond the Philippines), it did indeed occur, and followed an expansion into Taiwan from South China, as one archaeological model predicts (Bellwood 1997). However, we must be careful what we mean by the term "Neolithic", since the archaeological record for most of ISEA primarily indicates the appearance of various novel ceramics, and provides little or no evidence for large changes in the subsistence base. The low level of settlement across ISEA at this time accords not with large-scale demic diffusion fuelled by rice agriculture, but with more with archaeological views that stress the transition from grain cultivation to the root and arboreal crops that dominate agricultural systems in the western Pacific (Donohue and Denham 2010; Paz 2002). It is clearly parsimonious to conclude that these sea-faring settlers spoke Austronesian and spread their languages across ISEA, but they may have had rather little to do with either rice farming or arboriculture/vegeculture (aspects of which originated much earlier, in part diffusing from Near Oceania (Barker and Richards 2013; Blench 2012).

The low scale of the migrations overall concurs with recent archaeological evaluations (e.g. Spriggs 2011), but contrasts sharply with the recent interpretation of Lipson et al. (2014). However, their assumption that aboriginal Taiwanese represent the source for ISEA, their use of

only three autosomal source clusters and their extremely recent age estimates for admixture times (within the last 2200 years) compromise their conclusions. Our analysis supports a scenario in which language shift played the major role, rather than large-scale population replacement (Donohue and Denham 2010, 2015).

The genetic situation further east seems to require a model where language was transmitted mostly horizontally across the north coast of New Guinea. Curiously, M7c3c (most or all probably belonging to the subclade M7c3c2 dating to ~2.6 ka) and some other putative "out-of-Taiwan" subclades (like B4b1) are detected at relatively high frequencies in Eastern Micronesia/Northwest Polynesia. These lineages may have been carried directly through Western Micronesia from the vicinity of the Philippines (Fitzpatrick and Callaghan 2013; Hung et al. 2011). This migration was, however, distinct from the primary spread of the Austronesian languages into the Pacific, and would be expected to have affected mainly the Marianas.

Otherwise, whilst languages may have moved alongside other lineages integrated within ISEA, "out-of-Taiwan" haplogroups are virtually undetected across the north coast of New Guinea, the Bismarck Archipelago or the Solomon Islands. Minor exceptions include 1.4 % M7c3c in the Admiralty Islands (Kayser et al. 2008a), <0.2 % in New Britain (Friedlaender et al. 2007) and two closely related whole-mtDNA M7c3c sequences (~2 %) in the Solomon Islands (Duggan et al. 2014). M7c3c sequences, all within M7c3c2, are also seen in Ontong Java, a Polynesian outlier in the north Solomons. M7c3c and the other probable "out-of-Taiwan" clades have not been detected in Vanuatu, Fiji or Samoa, despite very extensive sampling.

Most of the present-day diversity in Near and Remote Oceania was established in New Guinea by ~10 ka (Soares et al. 2011), a fraction of which was carried by Austronesian speakers into the Remote Pacific. Powerful, long-established spheres of interaction may have facilitated the spread of the Austronesian languages in the south (Bulbeck 2008; Terrell and Welsch 1997; Torrence and Swadling 2008). They may thus have spread stepwise from the north and west via small-scale interactions and waves of acculturation. There appears to have been no "Austronesian farming-dispersal" in any meaningful sense across ISEA—early Austronesian speakers were more likely fisher–foragers—opening up the discussion to a range of innovative archaeological and linguistic models (Barker and Richards 2013; Donohue and Denham 2010, 2015). As both archaeologists and linguists have suggested, alluding to the spread of the early Metal Age in Europe, it may be that what began to spread across ISEA around 4000 years ago was primarily a new way of thinking—the adoption of a new ideology and perhaps even a new religion (Blench 2012; Spriggs 2011).

Acknowledgments This work was supported by FCT, the Portuguese Foundation for Science and Technology, through the project PTDC/IVC-ANT/4917/2012. PS is supported by FCT, European Social Fund, Programa Operacional Potencial Humano and the FCT Investigator Programme and acknowledges FCT/MEC for support to CBMA through Portuguese funds (PIDDAC)—PEst-OE/BIA/UI4050/2014. PS was also supported through the course of this work by a Marie Curie Early Stage Training Grant and FCT grant SFRH/BPD/64233/2009. MBR thanks the British Academy for financial support through project LRG-42440 and PS and MBR also thank the de Laszlo Foundation and the British Academy (project BARDA-48208). IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by FCT, the Portuguese Foundation for Science and Technology.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants All procedures performed in studies involving human participants were approved by the University of Leeds, Faculty of Biological Sciences Ethics Committee, and that of the University of Huddersfield, School of Applied Sciences.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen CH, Chen J, Chen YT, Chu J, Cutiongco-de La Paz EMC, De Ungria MCA, Delfin FC, Edo J, Fuchareon S, Ghang H, Gojobori T, Han J, Ho SF, Hoh BP, Huang W, Inoko H, Jha P, Jinam TA, Jin L, Jung J, Kangwanpong D, Kampuansai J, Kennedy GC, Khurana P, Kim HL, Kim K, Kim S, Kim WY, Kimm K, Kimura R, Koike T, Kulawonganunchai S, Kumar V, Lai PS, Lee JY, Lee S, Liu ET, Majumder PP, Mandapati KK, Marzuki S, Mitchell W, Mukerji M, Naritomi K, Ngamphiw C, Niikawa N, Nishida N, Oh B, Oh S, Ohashi J, Oka A, Ong R, Padilla CD, Palittapongarnpim P, Perdigon HB, Phipps ME, Png E, Sakaki Y, Salvador JM, Sandraling Y, Scaria V, Seielstad M, Sidek MR, Sinha A, Srikummool M, Sudoyo H, Sugano S, Suryadi H, Suzuki Y, Tabbada KA, Tan A, Tokunaga K, Tongsima S, Villamor LP, Wang E, Wang Y, Wang H, Wu JY, Xiao H, Xu S, Yang JO, Shugart YY, Yoo HS, Yuan W, Zhao G, Zilfalil BA (2009) Mapping human genetic diversity in Asia. Science 326:1541–1545

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655–1664

Anderson A (2005) Crossing the Luzon Strait: archaeological chronology in the Batanes Islands, Philippines and the regional sequence of Neolithic dispersal. J Austron Stud 1:25–44

Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, Choi Y, Van Duijn K, Vermeulen M, Brauer S, Decorte R, Poetsch M, Von Wurmb-Schwark N, De Knijff P, Labuda D, Vézina H, Knoblauch H, Lessig R, Roewer L, Ploski R, Dobosz T, Henke L, Henke J, Furtado MR, Kayser M (2010) Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am J Hum Genet 87:341–353

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Barker G, Richards MB (2013) Foraging-farming transitions in Island Southeast Asia. J Archaeol Method Theory 20:256–280

Bedford S, Spriggs M, Regenvanu R (2006) The Teouma Lapita site and the early human settlement of the Pacific Islands. Antiquity 80:812–828

Bellwood P (1997) Prehistory of the Indo-Malaysian archipelago. ANU E Press, Canberra

Bird MI, Austin WEN, Wurster CM, Fifield LK, Mojtahid M, Sargeant C (2010) Punctuated eustatic sea-level rise in the early mid-Holocene. Geology 38:803–806. doi:10.1130/g31066.1

Blanchon P, Shaw J (1995) Reef drowning during the last deglaciation: evidence for catastrophic sea-level rise and ice-sheet collapse. Geology 23:4–8

Blench R (2012) Almost everything you believed about the Austronesians isn't true. Crossing Borders: Selected Papers from the 13th International Conference of the European Association of Southeast Asian Archaeologists 2012. National University of Singapore Press, pp 122–142

Blust R (1976) Austronesian culture history: some linguistic inferences and their relations to the archaeological record. World Archaeol 8:19–43

Blust R (1995) The prehistory of the Austronesian-speaking peoples: a view from language. J World Prehist 9:453–510

Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, Jane Adler C, Richards SM, Sarkissian CD, Ganslmeier R, Friederich S, Dresely V, Van Oven M, Kenyon R, Van Der Hoek MB, Korlach J, Luong K, Ho SYW, Quintana-Murci L, Behar DM, Meller H, Alt KW, Cooper A, Adhikarla S, Ganesh Prasad AK, Pitchappan R, Varatharajan Santhakumari A, Balanovska E, Balanovsky O, Bertranpetit J, Comas D, Martínez-Cruz B, Melé M, Clarke AC, Matisoo-Smith EA, Dulik MC, Gaieski JB, Owings AC, Schurr TG, Vilar MG, Hobbs A, Soodyall H, Javed A, Parida L, Platt DE, Royyuru AK, Jin L, Li S, Kaplan ME, Merchant NC, John Mitchell R, Renfrew C, Lacerda DR, Santos FR, Soria Hernanz DF, Spencer Wells R, Swamikrishnan P, Tyler-Smith C, Paulo Vieira P, Ziegle JS (2013) Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nat Commun 4:1764

Bulbeck D (2008) An integrated perspective on the Austronesian diaspora: the switch from cereal agriculture to maritime foraging in the colonisation of Island Southeast Asia. Aust Archaeol 67:31–52

Bulbeck D (2011) Biological and cultural evolution in the population and culture history of Homo sapiens in Malaya. In: Enfield N (ed) Dynamics of human diversity: the case of mainland Southeast Asia. Pacific Linguistics, Canberra, pp 207–255

Capelli C, Wilson JF, Richards M, Stumpf MPH, Gratrix F, Oppenheimer S, Underhill P, Pascali VL, Ko TM, Goldstein DB (2001) A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. Am J Hum Genet 68:432–443

Clark GR, Anderson A (2009) The early prehistory of Fiji. ANU E Press, Canberra

Delfin F, Myles S, Choi Y, Hughes D, Illek R, Van Oven M, Pakendorf B, Kayser M, Stoneking M (2012) Bridging Near and Remote Oceania: mtDNA and NRY variation in the Solomon Islands. Mol Biol Evol 29:545–564

Diamond J, Bellwood P (2003) Farmers and their languages: the first expansions. Science 300:597–603

Donohue M, Denham T (2010) Farming and language in Island Southeast Asia: reframing Austronesian history. Curr Anthropol 51:223–256

Donohue M, Denham T (2015) Becoming Austronesian: mechanisms of language dispersal across southern Island Southeast Asia. In: Gil D, McWhorter J (eds) Austronesian Undressed. Pacific Linguistics, Canberra (in press)

Drummond AJ, Rambaut A (2007) BEAST: bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185–1192

Duggan AT, Evans B, Friedlaender FR, Friedlaender JS, Koki G, Merriwether DA, Kayser M, Stoneking M (2014) Maternal history of oceania from complete mtDNA genomes: contrasting ancient diversity with recent homogenization due to the Austronesian expansion. Am J Hum Genet 94:721–733

Fitzpatrick SM, Callaghan RT (2013) Estimating trajectories of colonisation to the Mariana islands, western Pacific. Antiquity 87:840–853

Forster P, Harding R, Torroni A, Bandelt H-J (1996) Origin and evolution of native American mtDNA variation: a reappraisal. Am J Hum Genet 59:935–945

Friedlaender JS, Friedlaender FR, Hodgson JA, Stoltz M, Koki G, Horvat G, Zhadanov S, Schurr TG, Merriwether DA (2007) Melanesian mtDNA complexity. PLoS One 2:e248

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo JH, Koki G, Hodgson JA, Merriwether DA, Weber JL (2008) The genetic structure of Pacific Islanders. PLoS Genet 4:0173–0190

Fu Q, Meyer M, Gao X, Stenzel U, Burbano HA, Kelso J, Pääbo S (2013a) DNA analysis of an early modern human from Tianyuan Cave, China. Proc Natl Acad Sci USA 110:2223–2227

Fu Q, Mittnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemsch L, Schmitz R, Burger J, Ronchitelli AM, Martini F, Cremonesi RG, Svoboda J, Bauer P, Caramelli D, Castellano S, Reich D, Pääbo S, Krause J (2013b) A revised timescale for human evolution based on ancient mitochondrial genomes. Curr Biol 23:553–559

Gusmão L, Sánchez-Diz P, Calafell F, Martín P, Alonso CA, Álvarez-Fernández F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML, Builes JJ, Capilla J, Carvalho M, Castillo C, Catanesi CI, Corach D, Di Lonardo AM, Espinheira R, Fagundes De Carvalho E, Farfán MJ, Figueiredo HP, Gomes I, Lojo MM, Marino M, Pinheiro MF, Pontes ML, Prieto V, Ramos-Luis E, Riancho JA, Souza Góes AC, Santapa OA, Sumita DR, Vallejo G, Vidal Rioja L, Vide MC, Vieira Da Silva CI, Whittle MR, Zabala W, Zarrabeitia MT, Alonso A, Carracedo A, Amorim A (2005) Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26:520–528

Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S, Richards M (2007) A mitochondrial stratigraphy for Island Southeast Asia. Am J Hum Genet 80:29–43

Ho SYW, Kolokotronis SO, Allaby RG (2007) Elevated substitution rates estimated from ancient DNA sequences. Biol Lett 3:702–705

Hudjashov G, Kivisild T, Underhill PA, Endicott P, Sanchez JJ, Lin AA, Shen P, Oefner P, Renfrew C, Villems R, Forster P (2007) Revealing the prehistoric settlement of Australia by Y

chromosome and mtDNA analysis. Proc Natl Acad Sci USA 104:8726–8730
Hung H-C, Carson MT, Bellwood P, Campos FZ, Piper PJ, Dizon E, Bolunia MJ, Oxenham M, Chi Z (2011) The first settlement of Remote Oceania: the Philippines to the Marianas: supplementary information on radiocarbon dating of the Nagsabaran site. Antiquity 85:909–926
Hunt CO, Gilbertson DD, Rushworth G (2007) Modern humans in Sarawak, Malaysian Borneo, during Oxygen Isotope Stage 3: palaeoenvironmental evidence from the Great Cave of Niah. J Archaeol Sci 34:1953–1969
Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey S, Lansing JS, Hammer MF (2010) Major east-west division underlies Y chromosome stratification across Indonesia. Mol Biol Evol 27:1833–1844
Kayser M, Brauer S, Weiss G, Underhill P, Roewer L, Schiefenhövel W, Stoneking M (2000) Melanesian origin of Polynesian Y chromosomes. Curr Biol 10:1237–1246
Kayser M, Choi Y, Van Oven M, Mona S, Brauer S, Trent RJ, Suarkia D, Schiefenhövel W, Stoneking M (2008a) The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Mol Biol Evol 25:1362–1374
Kayser M, Lao O, Saar K, Brauer S, Wang X, Nürnberg P, Trent R, Stoneking M (2008b) Genome-wide analysis indicates more Asian than Melanesian ancestry of Polynesians. Am J Hum Genet 82:194–198
Ko AMS, Chen CY, Fu Q, Delfin F, Li M, Chiu HL, Stoneking M, Ko YC (2014) Early Austronesians: into and out of Taiwan. Am J Hum Genet 94:426–436
Lipson M, Loh PR, Patterson N, Moorjani P, Ko YC, Stoneking M, Berger B, Reich D (2014) Reconstructing Austronesian population history in Island Southeast Asia. Nat Commun 5:4689
Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt H-J, Oppenheimer S, Torroni A, Richards M (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034–1036
Manguin P-Y, Mani A, Wade G (2011) Early interactions between South and Southeast Asia. Institute of Southeast Asian Studies, Singapure
Mellars P, Gori KC, Carr M, Soares PA, Richards MB (2013) Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proc Natl Acad Sci USA 110:10699–10704
Melton T, Peterson R, Redd AJ, Saha N, Sofro ASM, Martinson J, Stoneking M (1995) Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet 57:403–414
Ngamphiw C, Assawamakin A, Xu S, Shaw PJ, Yang JO, Ghang H, Bhak J, Liu E, Tongsima S, Consortium HP-AS (2011) PanSNPdb: the Pan-Asian SNP genotyping database. PLoS One 6:e21451
Oppenheimer S (1998) Eden in the east: the drowned continent of Southeast Asia. Phoenix, London
Oppenheimer S (2004) The 'Express Train from Taiwan to Polynesia': on the congruence of proxy lines of evidence. World Archaeol 36:591–600
Oppenheimer S, Richards M (2001) Fast trains, slow boats, and the ancestry of the Polynesian islanders. Sci Prog 84:157–181
Paz V (2002) Island Southeast Asia: spread or friction zone. In: Bellwood P, Renfrew C (eds) Examining the farming/language dispersal hypothesis. MacDonald Institute for Archaeological Research, Cambridge, pp 275–285
Pelejero C, Kienast M, Wang L, Grimalt JO (1999) The flooding of Sundaland during the last deglaciation: imprints in hemipelagic sediments from the southern South China Sea. Earth Planet Sci Lett 171:661–671
Pereira L, Silva NM, Franco-Duarte R, Fernandes V, Pereira JB, Costa MD, Martins H, Soares P, Behar DM, Richards MB, Macaulay V (2010) Population expansion in the North African Late Pleistocene signalled by mitochondrial DNA haplogroup U6. BMC Evol Biol 10:390
Pierron D, Razafindrazaka H, Pagani L, Ricaut FX, Antao T, Capredon M, Sambo C, Radimilahy C, Rakotoarisoa JA, Blench RM, Letellier T, Kivisild T (2014) Genome-wide evidence of Austronesian-Bantu admixture and cultural reversion in a hunter-gatherer group of Madagascar. Proc Natl Acad Sci USA 111:936–941
Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559–575
Redd AJ, Takezaki N, Sherry ST, McGarvey ST, Sofro ASM, Stoneking M (1995) Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. Mol Biol Evol 12:604–615
Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozzari R, Torroni A, Bandelt H-J (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251–1276
Rieth TM, Hunt TL (2008) A radiocarbon chronology for Sāmoan prehistory. J Archaeol Sci 35:1901–1927
Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, Soares P (2013) The first modern human dispersals across Africa. PLoS One 8:e80031
Ross M (2005) The Batanic languages in relation to the early history of the Malayo-Polynesian subgroup of Austronesian. J Austron Stud 1:P1–P24
Ross M (2009) Proto Austronesian verbal morphology: a reappraisal. In: Adelaar A, Pawley A (eds) Austronesian historical linguistics and culture history: a festschrift for Robert Blust. Pacific Linguistics, Canberra, pp 295–326
Saillard J, Forster P, Lynnerup N, Bandelt H-J, Norby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718–726
Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, Macaulay V, Bulbeck D, Oppenheimer S, Lin M, Richards MB (2008) Climate change and postglacial human dispersals in Southeast Asia. Mol Biol Evol 25:1209–1218
Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740–759
Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke DJ, Loo JH, Thomson N, Denham T, Donohue M, Macaulay V, Lin M, Oppenheimer S, Richards MB (2011) Ancient voyaging and Polynesian origins. Am J Hum Genet 88:239–247
Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, Costa MD, Musilová E, Macaulay V, Richards MB, Černý V, Pereira L (2012) The expansion of mtDNA haplogroup L3 within and out of Africa. Mol Biol Evol 29:915–927
Solheim WG (2006) Archaeology and culture in Southeast Asia: unraveling the Nusantao. UP Press, Philippines

Spriggs M (2003) Chronology of the Neolithic transition in Island Southeast Asia and the Western Pacific: a view from 2003. Rev Archaeol 24:57–80

Spriggs M (2007) The Neolithic and Austronesian expansion within Island Southeast Asia and into the Pacific. In: Chiu S, Sand C (eds) From Southeast Asia to the Pacific: Archaeological perspectives on the Austronesian expansion and the Lapita cultural complex. Academica Sinica, Taipei, pp 104–125

Spriggs M (2011) Archaeology and the Austronesian expansion: where are we now? Antiquity 85:510–528

Summerhayes GR, Leavesley M, Fairbairn A, Mandui H, Field J, Ford A, Fullagar R (2010) Human adaptation and plant use in highland New Guinea 49,000 to 44,000 years ago. Science 330:78–81

Sykes B, Leiboff A, Low-Beer J, Tetzner S, Richards M (1995) The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am J Hum Genet 57:1463–1475

Terrell JE, Welsch RL (1997) Lapita and the temporal geography of prehistory. Antiquity 71:548–572

Torrence R, Swadling P (2008) Social networks and the spread of Lapita. Antiquity 82:600–616

Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, Simionati B, Valle G, Richards M, Macaulay V, Scozzari R (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? Am J Hum Genet 69:1348–1356

Trejaut JA, Kivisild T, Jun HL, Chien LL, Chun LH, Chia JH, Zheng YL, Lin M (2005) Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. PLoS Biol 3:e376

Trejaut JA, Poloni ES, Yen JC, Lai YH, Loo JH, Lee CL, He CL, Lin M (2014) Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. BMC Genetics 15:77

Tumonggor MK, Karafet TM, Hallmark B, Lansing JS, Sudoyo H, Hammer MF, Cox MP (2013) The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. J Hum Genet 58:165–173

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonné-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ (2000) Y chromosome sequence variation and the history of human populations. Nat Genet 26:358–361

Verdu P, Pemberton TJ, Laurent R, Kemp BM, Gonzalez-Oliver A, Gorodezky C, Hughes CE, Shattuck MR, Petzelt B, Mitchell J, Harry H, William T, Worl R, Cybulski JS, Rosenberg NA, Malhi RS (2014) Patterns of admixture and population structure in native populations of Northwest North America. PLoS Genet 10:e1004530

Wollstein A, Lao O, Becker C, Brauer S, Trent RJ, Nürnberg P, Stoneking M, Kayser M (2010) Demographic history of Oceania inferred from genome-wide data. Curr Biol 20:1983–1992

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tournev I, Feldman MW, Kalaydjieva L (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74:50–61

PAPER II

Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia

Brandão A, *Eng KK, Rito T, Cavadas B, Bulbeck D, Gandini F, Pala M, Mormina M, Hudson B, White J, Ko T-M, Saidin M, Zafarina Z, Oppenheimer S, Richards MB, Pereira L, Soares P. Quantifying the* legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia. Hum Genet (Epub ahead of print). DOI 10.1007/s00439-016-1640-3

Hum Genet
DOI 10.1007/s00439-016-1640-3

ORIGINAL INVESTIGATION

Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia

Andreia Brandão[1,2,3,4] · Khen Khong Eng[5,6] · Teresa Rito[1,7,8] · Bruno Cavadas[1,2] · David Bulbeck[9] · Francesca Gandini[3] · Maria Pala[3] · Maru Mormina[5,10] · Bob Hudson[11] · Joyce White[12] · Tsang-Ming Ko[13] · Mokhtar Saidin[6] · Zainuddin Zafarina[14,15] · Stephen Oppenheimer[16] · Martin B. Richards[3,5] · Luísa Pereira[1,2,17] · Pedro Soares[1,2,5,18]

Received: 20 November 2015 / Accepted: 21 January 2016

Abstract There has been a long-standing debate concerning the extent to which the spread of Neolithic ceramics and Malay-Polynesian languages in Island Southeast Asia (ISEA) were coupled to an agriculturally driven demic dispersal out of Taiwan 4000 years ago (4 ka). We previously addressed this question using founder analysis of mitochondrial DNA (mtDNA) control-region sequences to identify major lineage clusters most likely to have dispersed from Taiwan into ISEA, proposing that the dispersal had a relatively minor impact on the extant genetic structure of ISEA, and that the role of agriculture in the expansion of the Austronesian languages was therefore likely to have been correspondingly minor. Here we test these conclusions by sequencing whole mtDNAs from across Taiwan and ISEA, using their higher chronological precision to resolve the overall proportion that participated in the "out-of-Taiwan" mid-Holocene dispersal as opposed to earlier, postglacial expansions in the Early Holocene. We show that, in total, about 20 % of mtDNA lineages in the modern ISEA pool result from the "out-of-Taiwan" dispersal, with most of the remainder signifying earlier processes, mainly due to sea-level rises after the Last Glacial Maximum. Notably, we show that every one of these founder clusters previously entered Taiwan from China, 6–7 ka, where rice-farming originated, and remained distinct from the indigenous Taiwanese population until after the subsequent dispersal into ISEA.

M. B. Richards, L. Pereira and P. Soares contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-016-1640-3) contains supplementary material, which is available to authorized users.

✉ Martin B. Richards
m.b.richards@hud.ac.uk

1 IPATIMUP (Institute of Molecular Pathology and Immunology of the University of Porto), Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

2 i3S (Instituto de Investigação e Inovação em Saúde, Universidade do Porto), 4200 Porto, Portugal

3 Department of Biological Sciences, School of Applied Sciences, University of Huddersfield Queensgate, Huddersfield HD1 3DH, UK

4 ICBAS (Instituto Ciências Biomédicas Abel Salazar), Universidade do Porto, Rua de Jorge Viterbo Ferreira n.º 228, 4050-313 Porto, Portugal

5 Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, UK

6 Centre for Global Archaeological Research, Universiti Sains Malaysia, 11800 Penang, Malaysia

7 Life and Health Sciences Research Institute (ICVS), School of Health Sciences, University of Minho, Braga, Portugal

8 ICVS/3B's-PT Government Associate Laboratory, Braga/Guimarães, Portugal

9 Department of Archaeology and Natural History, College of Asia and the Pacific, The Australian National University, Acton ACT, Canberra 2601, Australia

10 Department of Applied Social Studies, University of Winchester, Sparkford Road, Winchester SO22 4NR, UK

11 Archaeology Department, University of Sydney, New South Wales 2006, Australia

12 Department of Anthropology, University of Pennsylvania Museum, 3260 South St., Philadelphia, USA

13 Department of Obstetrics and Gynecology, National Taiwan University, Roosevelt Rd., Taipei 10617, Taiwan

14 Malaysian Institute of Pharmaceuticals and Nutraceuticals Malaysia, National Institutes of Biotechnology Malaysia, Penang, Malaysia

Published online: 13 February 2016

Introduction

Southeast Asia (SEA) harbours a rich variety of human populations with contrasting patterns of diversity seen in their ethnic cultures, languages, physical appearance and genetic heritage. The population history of this region was traditionally framed in terms of two distinct major prehistoric population movements. The first settlers, described as "Australo-Melanesian" people, arrived around 50–60 ka (thousand years ago) (Barker et al. 2007; Soares et al. 2009), and were the ancestors of several "Australoid" populations found in SEA, New Guinea and Australia (Bellwood 1995; Barker et al. 2007). The second migration occurred during the mid-Holocene (5–4 ka) and involved a large-scale demic expansion of rice agriculturalists starting in South China ~6 ka, which spread in two directions, one towards Mainland Southeast Asia (MSEA), and the other, via Taiwan, to Island Southeast Asia (ISEA), Near and Remote Oceania, and Madagascar (Bellwood 1995, 2005; Gray et al. 2009). Proponents of this "two-layer" model (Bellwood and Dizon 2008), drawn essentially from historical linguistics and some archaeological data, argue that the South Chinese rice agriculturalists partly or largely replaced the previous inhabitants of the region, whilst spreading Austronesian languages in ISEA and Austroasiatic languages in MSEA (Benedict 1976; Bellwood 1995; Bellwood et al. 2006). It is, however, possible that ISEA received direct influence from both of these hypothetical Neolithic migrations, as suggested by Anderson (2005), taking into consideration both archaeological and linguistic evidence. Anderson (2005) offered a more comprehensive view of the Neolithic spread in the region, suggesting that it most likely followed a reticulate pattern, and not a linear expansion model. He proposed the existence of two Neolithic movements from different sources: an earlier minor one ~4.5 ka from MSEA ("Neolithic I"), related to the spread of Austroasiatic languages and basket or cord-marked ceramics, into the Malay Peninsula and Borneo; and a second, major wave ("Neolithic II"), encompassing the hypothetical "out-of-Taiwan" migration (Bellwood and Dizon 2005, 2008). Our recent genetic work supports this view (Soares et al. 2016) but emphasizes that both mid-Holocene expansions were due to small-scale migrations.

Our genetic evidence suggests that other demographic events also contributed to current population structure in SEA, especially as a consequence of the massive climatic changes that occurred at the end of the Last Glacial Maximum (LGM). In the Late Pleistocene, ~20 ka, global sea levels were ~130 m below present-day levels, MSEA and Western ISEA were interconnected by a vast continental landmass, called Sundaland (Barker and Richards 2013), that facilitated early human dispersals through the region (Bird et al. 2005). After the LGM, rapid episodes of sea-level rises at ~14.5, 11.5 and 7.5 ka flooded about half of the land area of Sundaland, with a concomitant doubling of the length of the coastline (Oppenheimer 1998; Bird et al. 2005; Soares et al. 2008). Taking into consideration the past climatic changes in SEA, and the pressure suffered from the flooding of large areas of the landscape, some authors have suggested that these episodes triggered massive migratory events in the region (Oppenheimer 1998; Solheim 2006; Soares et al. 2008). Thus the dispersals across SEA could have resulted from movement and expansion of indigenous Southeast Asian people, possibly reflected in the increase in sites across ISEA at the end of the Pleistocene (O'Connor and Bulbeck 2014). Following this premise, Solheim's Nusantao Maritime Trading and Communication Network hypothesis (NMTCN) (Solheim 2006) argues that Southeast Asian natives, regardless of language, developed a highly maritime-oriented culture as a result of the changes in the climate and landscape in the region which promoted successful exchange systems between populations in the region for the past 10 ka. The cultural and linguistic similarities could then have been promoted through this wide-ranging trade and communication network.

Recent technological advances have led to the generation of huge amounts of new genetic data. Maternal, paternal and autosomal genetic markers have all been used to shed light on population migration history but genetic studies on SEA and the Pacific are often still framed within the two-layer model. For example, Friedlaender et al. (2008) suggested that the autosomal variation of Remote Pacific Islanders resulted almost solely from the mid-Holocene expansion of Austronesian-speaking Taiwanese, although their analysis did not include SEA populations. From a slightly different premise, Kayser et al. (2008) also argued that the Polynesian populations have clear maternal Asian ancestry, while Y chromosomes are mostly from New Guinean populations. On this view, Polynesian genetic make-up becomes the result of the intermarriage between Austronesian-speaking females carrying Asian mtDNA

[15] Human Identification Unit, School of Health Sciences, Health Campus, Universiti Sains Malaysia, Kelantan, Malaysia

[16] School of Anthropology, Institute of Human Sciences, The Pauling Centre, University of Oxford, 58a Banbury Road, Oxford OX2 6QS, UK

[17] Faculty of Medicine, University of Porto, Al. Prof. Hernâni Monteiro, 4200-319 Porto, Portugal

[18] Department of Biology, CBMA (Centre of Molecular and Environmental Biology), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

lineages (e.g. the mtDNA "Polynesian motif") with male Melanesians en route to the Pacific.

However, although the Polynesian motif (defining mtDNA haplogroup B4a1a1) is extremely frequent in the Remote Pacific, with ancestral lineages present equally in ISEA and Taiwanese aboriginals, this need not imply an "Austronesian dispersal". In fact, the Polynesian motif itself is absent in most of ISEA and not found further west of Wallace's line, except for southeast Borneo, and it has a coalescence time much greater than expected if it had emerged en route between Taiwan and the Pacific in the mid-Holocene (Soares et al. 2011). The molecular-clock evidence (strongly corroborated by archaeologically consistent estimates for the entry into Remote Oceania itself) rather suggests the ancestral lineage reached the Pacific in the Early Holocene (Soares et al. 2011), where it evolved into the Polynesian motif ~6–7 ka, probably in the Bismarck Archipelago, before expanding both east into the Remote Pacific and west back into ISEA.

In fact, an increasing number of studies in recent years have indicated that the simple two-layer expansion model does not capture the complexity of the demographic history in ISEA (Bulbeck 2008; Donohue and Denham 2010). Karafet et al. (2010), analysing patterns of Y chromosome variation (Y-SNPs), argued for a discontinuous four-phase colonization process with several population incursions in SEA, starting with the introduction of basal haplogroups with the first settlers, followed by Late Pleistocene/Early Holocene postglacial migrations from the mainland, the mid-Holocene "out-of-Taiwan", and a more recent migration in the historical era. Significantly, they suggest that only few paternal lineages are associated with the Austronesian dispersal, and that the other major lineages date to earlier population movements.

These results have been corroborated in other recent studies (Trejaut et al. 2014; Soares et al. 2016). In terms of mtDNA, although studies showed the existence of mtDNA lineages shared between Austronesian speakers of Formosan, Filipino and other ISEA populations (Trejaut et al. 2005; Tabbada et al. 2010), many have contradicted a demic "out-of-Taiwan" expansion due to the time frame (Trejaut et al. 2005; Hill et al. 2007; Soares et al. 2008, 2016). Moreover, some ISEA maternal mtDNA lineages did not trace back their origin to Taiwan, but instead arose within the ISEA region and spread toward Taiwan, probably because of climatic changes (Soares et al. 2008, 2016). For example, mtDNA haplogroup E underwent major expansions and dispersals in the Early to mid-Holocene, extending west into Malaysia, east into New Guinea and north into Taiwan, somewhere between 8 and 4 ka (Soares et al. 2008) (using the recalibrated mtDNA clock: Soares et al. 2009). Thus Taiwan appears to have been a recipient of haplogroup E lineages from the south, before the Austronesian dispersal, rather than being the major source of Holocene population migrations southwards across ISEA (as in the "out-of-Taiwan" model). Genome-wide analyses have independently supported the notion that Taiwan was, at least in part, the recipient of genetic input from ISEA, rather than the other way around (Abdulla et al. 2009).

Nevertheless, the genetic picture of SEA remains far from being fully understood. Recently, Soares et al. (2016) performed a founder analysis for ISEA that highlighted three major haplogroups representing the main signals in the analysis; two were postglacial or Early Holocene (haplogroups E and B4a1) and one was a mid-Holocene "out-of-Taiwan" marker (haplogroup M7c3c). Overall, the data, representing 30–40 % of all present-day mtDNA lineages, matched the Early Holocene period, implying that although migrations from Taiwan did occur in the mid- to late Holocene, the so-called Austronesian expansion was mainly a process of cultural diffusion and assimilation. The remaining mtDNA lineages, many displaying low frequencies, cannot be so clearly partitioned using a founder analysis based on HVS-I sequences (first hypervariable segment of the control region). Here, therefore, we analyse in detail the sequence variation of whole-mtDNA genomes ("mitogenomes") of these low frequency mtDNA lineages. These lineages have already been tentatively associated with various demographic events in SEA, including the first settlement (haplogroup F3, R9b), Early Holocene postglacial expansions (haplogroup R9c, N9a) and mid-Holocene dispersals from Taiwan (haplogroups B4c1, F1a4, B5b, Y2, B4b1 and D5) (Hill et al. 2006, 2007; Soares et al. 2016), potentially identifying the spread of Neolithic material culture. We previously analysed R9b with whole mtDNAs (Hill et al. 2006), but the subsequent increase in sampling, as well as a revision of the molecular clock (Soares et al. 2009) demand a reassessment of the phylogeography of the clade. A comprehensive study of these low-frequency haplogroups in Southeast Asia can complete the picture of both the main dispersal routes and the impact of dispersals on the population history in the region. Our study ranges across the vast geographic region of Taiwan, MSEA, ISEA and Near Oceania, in contrast to other recent studies (Tabbada et al. 2010; Loo et al. 2011; Ko et al. 2014), in which more limited geographic regions were targeted.

Methods

Population samples and whole-mtDNA sequence analysis

We selected the population samples used in this study on the basis of the information from mtDNA hypervariable segment I (HVS-I), which allowed the broad classification

of the samples into haplogroups. We selected 114 samples belonging to 10 haplogroups in the region of Southeast Asia (B4b1; B4c1; B5b, D5; F1a4; F3, N9a; R9b, R9c and Y2). We chose the samples to constitute a dataset representative of the genetic variability of the general population. We included 2 from China, 23 from Taiwan, 61 from MSEA (22 from Vietnam; one from Thailand; 31 from Peninsular Malaysia; 4 from Laos and three from Myanmar (Burma)); 26 from ISEA (12 from island of Borneo—three from Kota Kinabalu (Sabah, Malaysian state), five from Brunei, four from Palangkaraya (Indonesian province Central Kalimantan); 11 from other parts of Indonesia and three from the Philippines), and two from Micronesia (details in Table S1). The work was approved by the University of Huddersfield, SAS Ethics Committee.

For the whole-mtDNA genome sequencing, we followed the methodology and checking procedures reported in Pereira et al. (2010) and we analysed the sequences with BioEdit 7.0.4.1 (Hall 1999) and Sequencher 5.2.3 sequences analysis software (Gene Codes Corporation, Ann Arbour, MI, USA). We deposited 114 new whole-mtDNA sequences in GenBank (Accession Numbers KU521394-KU521507).

To obtain detailed phylogenetic reconstruction and precise age estimates of clades and times of expansion, we took comparative data from the literature. More specifically, we used 829 published whole-mtDNA genomes, which included 52 from MSEA, 197 from Taiwan, 173 from ISEA, and 407 from neighbouring regions (for more detailed information see Table S2).

Statistical analyses

To avoid any nomenclature conflicts, we followed the criterion of PhyloTree [mtDNA tree Build 16 (19 Feb 2014)] (van Oven and Kayser 2009). We disregarded the transition at 16,519 and the C-length polymorphisms in regions 16,180–16,193 and 309–315 in the analyses (Soares et al. 2009). We performed the classification of the variants with mtDNA GeneSyn (Pereira et al. 2009), and scored mutations in relation to the revised Cambridge reference sequence (rCRS) (Andrews et al. 1999).

We performed preliminary reduced-median network analyses (Bandelt et al. 1995), providing a suggested branching order for the trees, and used these to construct a putative most-parsimonious tree based on the relative mutation rates of the different positions (Soares et al. 2009).

For estimation of the coalescence times for specific clades in the phylogeny, we used the ρ statistic and maximum likelihood (ML). We used the ρ statistic with a mutation-rate estimate for the whole-mtDNA sequence of one transition in every 3624 years, corrected for purifying selection using the calculator developed by Soares et al. (2009), and a synonymous mutation rate of one substitution every 7884 years. We estimated standard errors as in Saillard et al. (2000). We also obtained ML estimates of branch lengths using PAML 4.7a (Yang 1997), assuming the HKY85 mutation model with gamma-distributed rates (approximated by a discrete distribution with 32 categories). We converted mutational distance in ML to time using the whole-mtDNA genome clock described above (Soares et al. 2009).

To access the demographic changes through time in SEA populations associated with the haplogroups studied, we obtained BSPs (Drummond et al. 2005; Fagundes et al. 2008) using BEAST (version 1.7.5) with a relaxed molecular clock (lognormal in distribution across branches and uncorrelated between them) using a mutation rate of 2.6186×10^{-8} mutations per site per year for the whole-mtDNA genome (Soares et al. 2012) and the HKY model of nucleotide substitutions with gamma-distributed rates, assuming a generation time of 25 years. In addition, we forced the larger subclades into monophyly to obtain a tree structure that was directly comparable with the remaining analyses (Fagundes et al. 2008). We visualised the plots with Tracer v1.3, and inferred the increment ratio by calculating the number of times that the effective population size increased during specific periods. For a broader overview of the geographic distribution patterns of the lineages, we constructed interpolation maps of spatial frequencies based on their HVS-I sequences (in the range of 16,051–16,400 bp), using the Kriging algorithm of Surfer 8 (Fig. S1).

We estimated founder ages for the main clades present in Taiwan, based on the whole mitogenomes. To minimize the impact of recurrent mutation and back-migrations, we selected founders on the strength of their source diversity, using an *f1* criterion, which stipulates that a sequence type is only considered a founder if it presents at least one derived branch in the source population (Richards et al. 2000). We calculated founder ages and confidence intervals as described before (Richards et al. 2000; Soares et al. 2012; Rito et al. 2013) and plotted the overall pattern at 200-years intervals. Since the current founder analysis methodology does not incorporate a time-dependent molecular clock, we made an approximation for the time scale under study. The mutation rate varied between 1 mutation per 2599 and 2679 years for the point estimates of the estimated founders, so we used an average value of 2639.

Results

General patterns of migration and expansion in Island Southeast Asia

For the phylogeographic analysis, we used 870 previously published and 114 newly sequenced mitogenomes

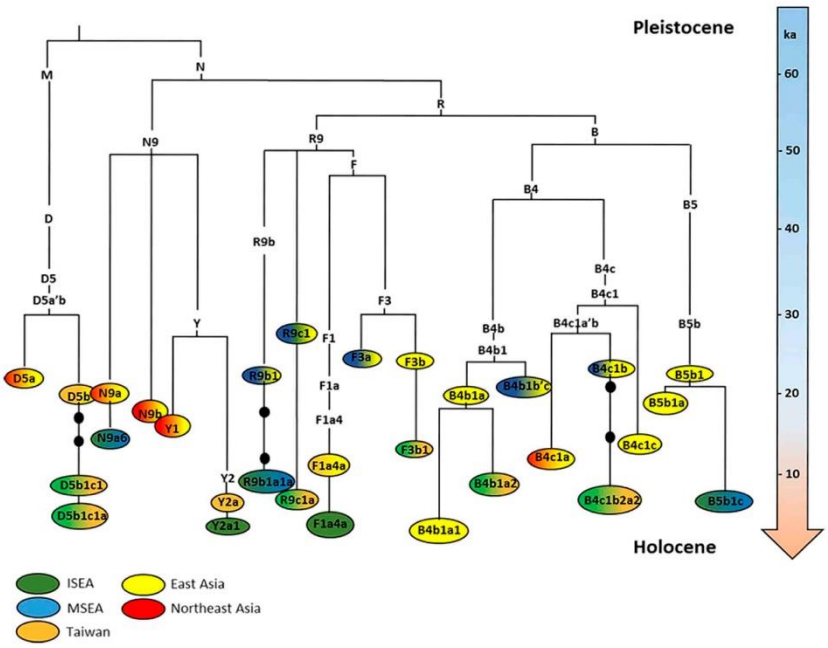


Fig. 1 Schematic tree of the subclades most representative in SEA belonging to haplogroups B4b1, B4c1, B5b, D5, F1a4, F3, N9a, R9b, R9c and Y2. The higher-frequency lineages B4a1a, E1, E2 are not shown in the figure, since they were analysed previously by Soares et al. (2016) Tree scaled using maximum likelihood and time-dependent molecular clock for whole-mtDNA genome (in ka). The *shading* represents the geographic distribution of the subclades. Details of age estimates are shown in Table 1

belonging to haplogroups R9b, R9c, F1a4, F3, F4b, B4b1, B4c1, B5b, N9, Y and D5 (Online Resource 1, namely Table S1, Table S2 and Supplementary Note 1, and Online Resource 2). Figure 1 shows an outline topology of the main subclades in East Asia and SEA for these haplogroups, scaled against the ML age estimates (for detail of ρ and ML age estimates, see Table S3). F4b, which entered Taiwan from China at the time of the Neolithic, but does not disperse further into ISEA, is not included. We can group the haplogroups into those with Early Holocene and those with mid-Holocene ancestry in ISEA (Figs. S2 and S3). The clades B4a1, E1, E2 (the higher-frequency lineages analysed previously (Soares et al. 2016), not shown in Fig. 1), F3b1, R9c1a, B5b1c and B4c1b2a2, corresponding to almost 27 % of all present-day mtDNA lineages in ISEA, most probably expanded within ISEA mainly between 10 and 7 ka, many of them reaching Taiwan at some point in the last 8 ka. Haplogroups M7c3c, Y2a1, B4b1a2, F1a4, D5b1c1 and M7b3, amounting to ~20 % overall in ISEA primarily show founder ages that indicate a mid-Holocene, potentially Neolithic entrance into this region, probably from a source in Taiwan (Table 1).

The tree topology for each clade reinforces these inferences. The first group (encompassing the higher frequency Early Holocene candidate haplogroups analysed previously by Soares et al. (2016), B4a1a, E1, E2, and the remaining lower-frequency haplogroups analysed in this study, F3b1, R9c1a, B5b1c and B4c1b2a2) displays long branches with a lack of branching nodes, indicating bottleneck/drift between, in general, 20–15 and 10 ka. More specifically, 15.1–9.9 ka from B4a1 to B4a1a (four mutations), 24.0–6.7 ka in B5b1 and B5b1c (seven mutations), 25–6 ka in F3b to F3b1 (six mutations), 14.5–8 ka from B4c1b2a to B4c1b2a2 (two mutations), 28.5–6 ka from R9c1 to R9c1a (eight mutations), and 24–8.3 and 12.7 ka from E to E2 (five mutations) and E1 (four mutations), respectively (Fig. S4). On the other hand, the mid-Holocene "out-of-Taiwan" clades (M7c3c, Y2a1, B4b1a2, F1a4, D5b1c1 and M7b3) show considerably shorter branch lengths separating the clade that expanded within current Austronesian-speaking populations and the ancestral point in continental Asia. The age estimate of the latter is usually within the 12–6 ka range: 11.7–4.3 ka from F1a4a to F1a4a1 (two mutations), 9.6–6.1 from Y2 to Y2a (one mutation), 17.7–9.3 ka in B4b1a + 207 to B4b1a2 (one mutation; but B4b1a2 exists in mainland China and Japan and is probably the founder clade, meaning that no mutation separates the Asian ancestor and the ISEA clade), 11.8–5.3 ka from M7c3 to M7c3c (one mutation) and 7.9–5.7 from M7b3 to M7b3a (one mutation) (Fig. S4).

A plausible explanation for this pattern may lie in the fact that the time interval before 8 ka was very likely a period of low effective population size in ISEA, with expansions largely restricted to the last 10 ka. The haplogroups that probably expanded in the Early Holocene postglacial period could therefore represent clades that were already present

Table 1 Age estimates using rho (ρ) and ML for major subclades in ISEA for haplogroups B4b1, B4c1, B5b, D5, F1a4, F3, N9a, R9b, R9c and Y2. Ages and 95 % confidence intervals (CI) in thousands of years

mtDNA lineage	*N*	PAML		Rho			
				Total		Synonymous	
		Age	95 % confidence interval	Age	95 % confidence interval	Age	95 % confidence interval
B4b1	129	25,100	[17,000–33,600]	22,600	[12,600–33,200]	32,600	[9400–55,900]
B4b1a2	89	9300	[6800–11,800]	8700	[6400–11,000]	9100	[5300–12,900]
B4c1b2a	56	14,500	[6400–23,000]	11,200	[3800–19,000]	15,500	[0–31,400]
B4c1b2a2	53	8000	[5600–10,500]	5800	[3500–8100]	7700	[1900–13,600]
B5b	89	29,900	[20,700–39,300]	34,300	[24,300–44,700]	43,500	[24,500–62,500]
B5b1	54	23,900	[13,900–34,300]	27,000	[15,800–38800]	44,700	[19,200–70,300]
B5b1a	15	19,000	[8,800–29,800]	20,500	[9300–32,400]	12,600	[0–25,800]
D5	174	33,300	[24,600–42,200]	34,500	[23,000–46,500]	35,600	[15,500–55,800]
D5b1c1	15	9100	[4000–14,400]	12,300	[4300–20,600]	21,000	[900–41,100]
D5b1c1a	11	6000	[0–13,800]	7400	[1200–13,700]	8600	[0–19,500]
D5b3	40	10,900	[5600–16,400]	9900	[2400–17,700]	17,700	[0–37,100]
F1a4	26	16,300	[7000–25,900]	18,600	[6500–31,400]	11,500	[0–26,100]
F1a4a	25	11,700	[3000–20,900]	10,600	[2600–19,000]	11,700	[0–26,800]
F1a4a1	23	4300	[1800–6800]	5200	[1500–9000]	3900	[800–7100]
F1a4a1a	15	3300	[1300–5300]	3500	[1500–5500]	3700	[1000–6400]
F3	88	31,700	[21,500–42,300]	37,900	[22,900–53,600]	35,300	[12,400–58,400]
F3a	20	26,600	[16,500–37,200]	31,500	[18,700–44,900]	26,400	[6000–46,800]
F3a1	16	16,600	[9000–24,500]	15,600	[9300–22,100]	13,300	[1300–25,300]
F3b	68	25,200	[15,400–35,400]	28,900	[13,600–45,100]	27,800	[3500–52,200]
F3b1	65	12,400	[5200–20,000]	12,000	[4700–19,700]	12,100	[0–25,100]
N9	254	50,600	[37,100–64,600]	38,500	[27,700–49,600]	36,100	[20,700–51,400]
N9a	127	20,000	[14,500–25,500]	17,500	[13,000–22,100]	18,400	[10,900–25,900]
N9a10	18	16,600	[11,000–22,400]	14,600	[8800–20,600]	14,900	[4600–25,200]
N9a10a	9	10,000	[4700–15,400]	9100	[4100–14,200]	14,000	[1600–26,400]
N9a10a1	6	6300	[200–12,500]	5200	[1200–9300]	6600	[0–14,300]
N9a6	45	14,800	[9900–19,800]	12,700	[7100–18,500]	9600	[2900–16,400]
R9b	45	38,700	[23,900–54,300]	32,800	[20,400–45,900]	31,400	[13,600–49,200]
R9b1a	32	18,600	[10,900–26,700]	20,700	[13,200–28,500]	23,900	[9800–38,100]
R9b1a1	18	11,600	[6000–17,300]	13,400	[6600–20,300]	9600	[3800–15,500]
R9b2	4	5700	[1300–10,200]	5200	[1600–8900]	5900	[0–12,600]
R9c1	45	28,500	[17,200–40,300]	25,200	[13,200–37,800]	33,100	[9600–56,600]
R9c1a	33	5900	[3700–8300]	5100	[2600–7700]	6900	[400–13,500]
Y	98	28,000	[16,100–40,500]	24,250	[14,000–34,900]	31,200	[11,600–50,900]
Y2	50	9600	[5000–14,400]	9250	[3100–15,600]	8800	[0–20,100]
Y2a	43	6100	[3200–9100]	6700	[2000–11,500]	8400	[0–21,400]
Y2a1	36	4100	[2300–5900]	4500	[2400–6600]	1500	[400–2700]

in ISEA by the time of the flooding of the Sunda shelf, and were maintained during the low effective population size period, and later expanded ~10–7 ka. On the other hand, the candidate mid-Holocene "out-of-Taiwan" clades were present in mainland China by 10 ka and moved to Taiwan and ISEA in the last 7–4 ka; and they do not display similar long branches but only branches with one or two mutations linking them to the mainland ancestor.

This is also reflected in the BSPs (Bayesian skyline plots, Fig. S5), which display changes in effective population size over time. No increment in population size is observed previous to 10 ka except at ~60 ka, which reflects mainly the basal M, N and R "out-of-Africa" founder nodes and occurred outside ISEA (Macaulay et al. 2005). Most of the clades analysed in the BSP (except maybe for haplogroups E and F3b) were not present in ISEA before

Table 2 Peaks of population size through time as obtained from BSPs for all haplogroups examined in this study, and the overall data for ISEA

Data set	Peak (ka)	Range of increment (ka)	Increment ratio
Haplogroup			
N9	11.9	9.8–14.9	0.5
	6.8	3.8–7.7	1.8
B4b1	6.1	4.9–7.0	2
	0.8	0–2.0	6
B4c1	6.1	5.0–7.6	2.5
Y2	3.1	2.2–4.2	1.1
F3	7.4	5.2–10.3	0.1
	2.5	4.1–1.2	1.1
R9b	5.4	0.9–7.1	1.5
B5b	16.8	15.8–18.9	0.7
	3.9	3.3–5.0	7.9
D5	7.7	3.5–13.1	2.1
R9c	3.6	2.1–5.2	1.7
F1a4	2.4	1.7–4.0	0.3
Region			
ISEA	9.0	8.2–10.6	10.1
	2.6	2.3–4.1	76.9

20–15 ka. From about 10.5–8 ka, a 19-fold increment in the population size occurred, followed by a 70-fold increase in the mid-Holocene, starting ~4 ka (Table 2). The BSP containing the clades analysed here (as well as all B4a, M7 and E clades) indicates two time periods of population expansion that perfectly fit the time of expansions inferred from the phylogeographic analysis.

It is worth emphasizing that the above mentioned increments correspond to clade expansions, which do not represent their relative frequencies in the ISEA population—for example, the low frequency clades analysed here are actually oversampled in relation to the more common clades. Furthermore, more ancient basal clades from SEA, namely basal M clades (Hill et al. 2007), were not included, and these could well have expanded in both periods.

Founder mtDNAs in Taiwan

We separated founders entering Taiwan into lineages that had a clear mainland origin in the last 15 ka and those that probably expanded in ISEA and Taiwan in the Early Holocene period from a source in ISEA (Table S5 and Fig. 2a). The latter generated a single peak at just above 6 ka (line in black in Fig. 2b). The clades for which we postulated a mainland origin led to a pattern that showed two peaks, one at 7.6 ka and another at 11 ka. These could be, respectively, correlated with the entry of rice agriculturalists, in accordance with the "out-of-Taiwan" model, and to arrivals in the postglacial Early Holocene period (or simply marking the separation of continental China and Taiwan due to sea-level rises). We further performed the analysis by stipulating two points of migration, one at 6.5 ka (following the "out-of-Taiwan" model) and one at 11 ka, to allocate each clade probabilistically to a migration event.

Four of the clades, B5a2, B4a2, D5b3 and R9b1a2, were probably present in Taiwan for more than 10 ka. All the other clades analysed (B4b1a2, F1a4, M7c3, Y2a, N9a10, M7b3a and M7b1d3) showed a higher probability of an entry with the Neolithic material culture (Fig. 2a). An important feature is that all the clades that we detected as "out-of-Taiwan" in the phylogeographic analysis were present here as input from the mainland dating to the spread of the rice Neolithic. This indicates two things: firstly, it provides further support to identify these clades as "out-of-Taiwan" candidates; but it also suggests that (at least until the "out-of-Taiwan" migration into ISEA) there was some level of separation of the expanding population from the autochthonous population of Taiwan, since the more ancient clades in Taiwan, as well as the ones entering before this time from ISEA, do not show any evidence of playing a role in the "out-of-Taiwan" migration.

We have further studied the distribution of the clades in the Taiwanese aboriginal populations according to their ancestry (Fig. 2c). Clades from China that show an Early Holocene ancestry in Taiwan are much more frequent in the south and almost absent on the east coast of Taiwan (Fig. 2c.1). The clades that we infer to have entered Taiwan carried by rice-agriculturalists from South China are much more frequent in the northern tribes and east coast of Taiwan (Fig. 2c.2). The influx clades that are of Early Holocene age in ISEA are much more evenly distributed across the extant aboriginal groups of Taiwan, with a peak on the east coast and a lower frequency in the northern tribes (Fig. 2c.3).

Genetic ancestry of the "out-of-Taiwan" clades

Since all the aboriginal groups now carry some clades that were involved in the "out-of-Taiwan" event (Fig. 3), we checked the best matches between this "out-of-Taiwan" composition and relative frequencies of those clades in the Taiwanese aboriginal groups, in order to provide some insights into the origin of the expanding population from Taiwan. Although northern tribes are the only ones that contain haplogroup Y2a, they lack the major "out-of-Taiwan" clade, M7c3c but carry a very high frequency of M7b3, which is a minor clade in ISEA. Most probably, considering the drift that the Taiwanese aboriginal groups have undergone (due to their small size and isolation) in the last few thousand years, Y2a was probably lost due to drift

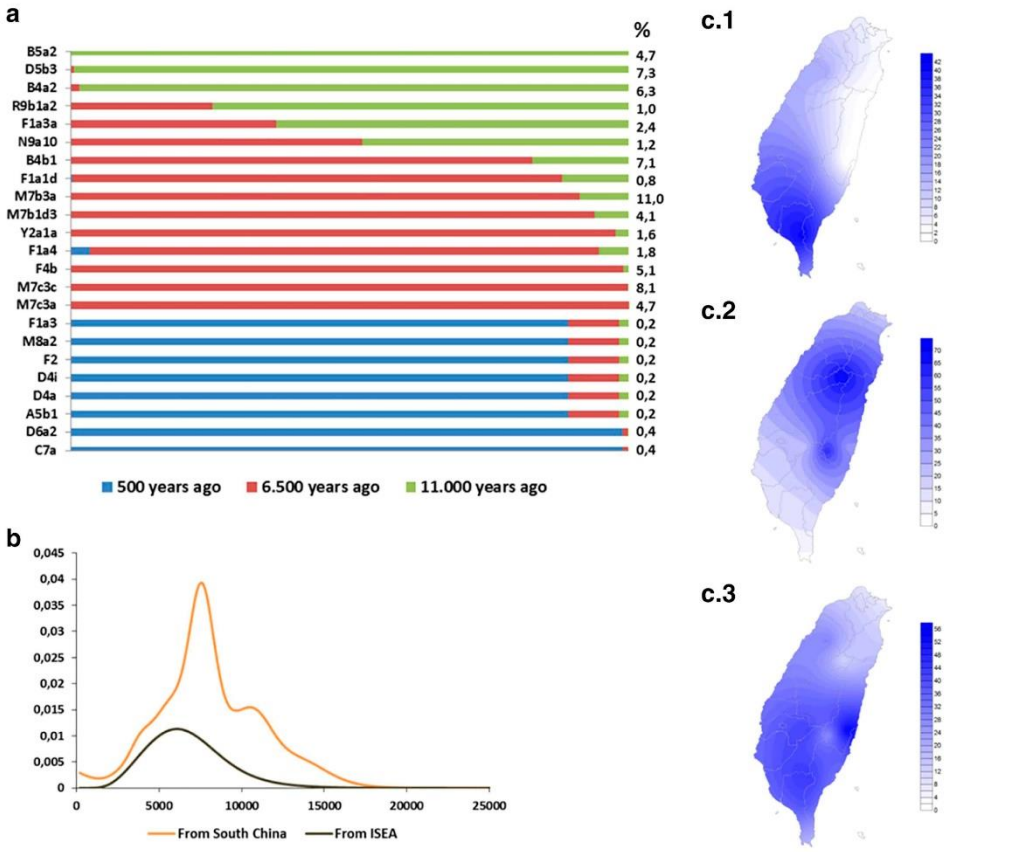


Fig. 2 Analysis of maternal genetic flow into Taiwan. **a** Probabilistic distribution of founders from mainland Asia, assuming three migrations, using *f1* criterion; **b** scan of migration time into Taiwan from South China (*orange line*) and ISEA (*black line*). **c** Frequency distribution maps of Taiwan based on HVS-I data: *c1* Pooled frequency of candidate postglacial mainland South China haplogroups (B5a2, B4a2, D5b3 and R9b1a2); *c2* Pooled frequency of candidate Neolithic South China haplogroups (B4b1a2, F1a4, M7c3, Y2a, F4b, N9a10, M7b3a and M7b1d3); *c3* Pooled frequency of candidate ISEA influx haplogroups (B4a1a, B5b1c, F3b1a, B4c1b2a2, E1 and E2). The map of Taiwan was adapted under the terms of the GNU Free Documentation License

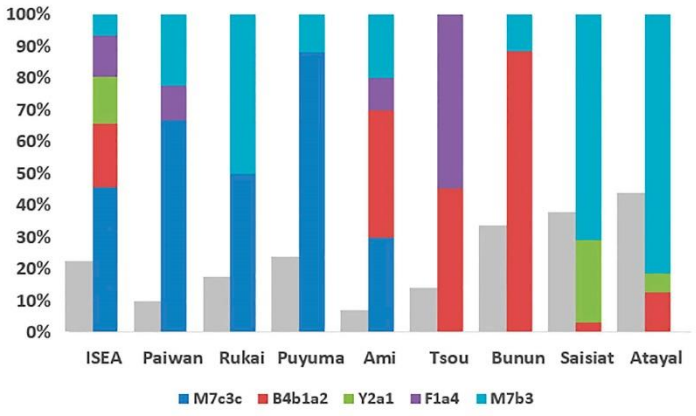


Fig. 3 Estimated contributions of Taiwanese "out-of-Taiwan" mtDNA lineages in the ISEA and Taiwanese aboriginals gene pool. The *grey bar* represents the overall frequency of those lineages in each population and the *second bar* represents the relative frequency of those haplogroups within each population

at other locations in Taiwan. The southern groups show the highest frequency of M7c3c, the major mid-Holocene "out-of-Taiwan" marker, but as well as Y2a they also lack B4b1, the second most common "out-of-Taiwan" clade in ISEA. B4b1 and F1a4 are more common in the central Taiwan tribes, but the overall composition that more closely matches the inferred group that dispersed into ISEA is that of the Ami of the east coast. They show a relatively high frequency of M7c3c and B4b1, the two major "out-of-Taiwan" clades in ISEA, and also carry M7b3 and F1a4. Curiously, the overall frequency of clades entering Taiwan at 6.5 ka in the founder analysis was higher for the northern tribes (such as Saisiat), as well as the Ami. Therefore the genetic evidence is concordant with the linguistic evidence suggesting that the Ami language is the closest Formosan candidate to the Malayo-Polynesian branch of the Austronesian family (Ross 2005).

Discussion

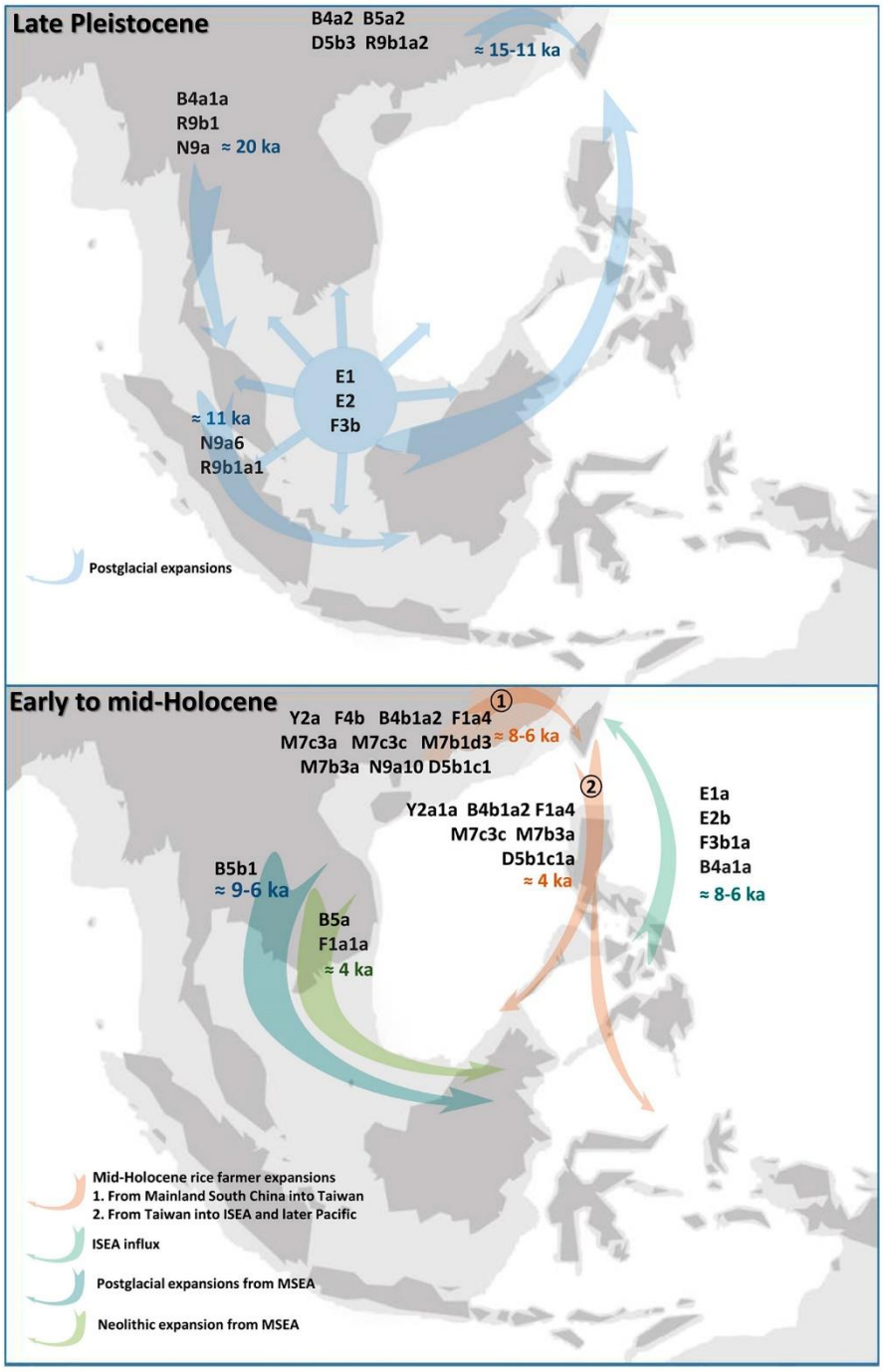
Vigorous debate continues among archaeologists, linguists and geneticists as to which major population movements shaped the demographic history of Southeast Asia. Here we tested a number of human mtDNA clades, individually present at low frequency, but together amounting to a quarter of the ISEA maternal component, and which had previously been suggested as representing various different stages of the genetic history of the region. We have shown that the phylogeography of these low-frequency clades closely matches two major hypothetical events, each one already recognised from the more frequent clades: namely, postglacial expansions in the Early Holocene and an "out-of-Taiwan" dispersal in the mid-Holocene. This study therefore complements our earlier work on the more frequent B4a1a, E and M7 clades, which represent about a third of the mtDNA lineages in ISEA.

At the end of the LGM, SEA underwent a major environmental change, with rising sea levels flooding many low-lying areas of the Sunda shelf, eventually creating the modern coastline (Bird et al. 2005). The loss of almost half of the land area and the environmental transformations, with the replacement by rainforest of the savannah and monsoon forest to which the Sunda populations had adapted, is thought to have caused enormous population displacements and cultural changes (Bird et al. 2005; Solheim 2006; Soares et al. 2008; Tumonggor et al. 2013). The effective population size is likely to have been low during this cataclysmic period. We would also expect that the later expanding populations would be composed of an unrepresentative sample of the maternal pool of the Sunda populations that existed from 20 to 10 ka. From our analysis, the clades that likely expanded in the Early Holocene postglacial have three basic characteristics: a split time with continental Asia of at least 15 ka; a lack of ancestral branching nodes between 15 and 8 ka; and, of course, an estimate of the age of expansion of the clade mostly between 10 and 7 ka. Haplogroups B4a1a, B5b1c, F3b1, B4c1b2a2, R9c1a and E fit this pattern well; and their BSPs also show lower effective population sizes until about 10 ka ago. Undoubtedly, many lineages from ISEA must have been lost during that period, potentially explaining the lack of ancestral lineages, such as the ancient DNA E1 lineage detected further north (Ko et al. 2014) when the overall pattern of haplogroup E strongly suggests ISEA as the point of origin and evolution for the clade (Soares et al. 2008) and the age estimate based on the general mtDNA clock and on ancient DNA calibration places the clade within ISEA much before the putative migration of rice agriculturists into Taiwan (Soares et al. 2016).

The observed two-way Early Holocene population expansions between East and Southeast Asia probably did not take place as a single monolithic event, but in multiple radiations from 10 ka onwards [cf. the "early train" hypothesis (Jinam et al. 2012)]. Eventually, some of the lineages reached Taiwan. The founder peak indicated a period ~ 6 ka, but ISEA lineages might have reached the island at several times in the postglacial period. The frequency patterns of these lineages in Taiwan clearly display a south-to-north cline as expected from lineages arriving from ISEA (Fig. 4).

A founder analysis from China into Taiwan showed that part of Taiwan's aboriginal maternal gene pool also included lineages that were present there for more than 10 ka (Chang 1989; Olsen and Miller-Antonio 1992). Some of them could be lineages that were present in the population of that part of the Asian continent before sea-level rise separated Taiwan from the mainland. Consistent with this, these ancient lineages are present throughout all the Taiwanese tribes.

In the last 7 ka, ISEA was not the only source of gene flow into Taiwan. Dispersals from South China also occurred in this time frame, and according to the prevailing "out-of-Taiwan" model these individuals were rice agriculturalists. The "out-of-Taiwan" model states that Neolithic farmers who had settled in Taiwan from South China around 7–6 ka, amongst whom the Austronesian languages arose, spread south after ~4.5 ka into the Philippines, Indo-Malaysia and into the Pacific, carrying with them the Neolithic package and the Proto-Malayo-Polynesian language (Ko et al. 2014). Here we see that B4b1a2, F1a4a, Y2a1 and the very minor clade D5b1c1 each show a strong signal of Taiwanese Neolithic ancestry in ISEA, forming—together with M7c3c and M7b3—the bulk of the "out-of-Taiwan" ancestry in ISEA (Fig. 4).


Late Pleistocene
B4a2 B5a2
D5b3 R9b1a2
≈ 15-11 ka
B4a1a
R9b1
N9a ≈ 20 ka
E1
E2
F3b
≈ 11 ka
N9a6
R9b1a1
Postglacial expansions
Early to mid-Holocene
①
Y2a F4b B4b1a2 F1a4
M7c3a M7c3c M7b1d3
M7b3a N9a10 D5b1c1
≈ 8-6 ka
②
Y2a1a B4b1a2 F1a4
M7c3c M7b3a
D5b1c1a
≈ 4 ka
E1a
E2b
F3b1a
B4a1a
≈ 8-6 ka
B5b1
≈ 9-6 ka
B5a
F1a1a
≈ 4 ka
Mid-Holocene rice farmer expansions
1. From Mainland South China into Taiwan
2. From Taiwan into ISEA and later Pacific
ISEA influx
Postglacial expansions from MSEA
Neolithic expansion from MSEA

◀**Fig. 4** Outline of maternal lineages involved in the main human migrations in the region of Southeast Asia and Taiwan. Includes those discussed here and also those described previously in Soares et al. (2016), including B5a and F1a1a, which were inferred to have dispersed from MSEA with the Neolithic. *Dark shading* represents the modern coastlines and the extent of Sundaland at the LGM is represented by the *light shading*. The map was obtained from the website http://www.outline-world-map.com

We can therefore refine our estimate of the fraction of Taiwanese Neolithic lineages present in ISEA today. It is theoretically possible that the frequency of mid-Holocene "out-of-Taiwan" founders into ISEA might be slightly underestimated as some of the founders that entered Taiwan from ISEA might have back-migrated to ISEA in the mid-Holocene. However, our results showed no evidence that pre-existing autochthonous lineages in Taiwan were involved in the mid-Holocene "out-of-Taiwan" migration, suggesting that there was a degree of isolation between those ancient lineages and the ones expanding from South China. Given this, and considering that we did not detect any putative mid-Holocene founders within the mitogenome tree of B4a1a, E1, E2 and F3b, it seems likely that these lineages, as the autochthonous lineages in Taiwan, were isolated from the putative "out-of-Taiwan" migrating population. Overall, our ISEA database (2117 samples from the Philippines and Indonesia) includes 19.5 % Neolithic lineages. The overall value is not very meaningful, however, because the estimate varies considerably from region to region. As one would expect from the inferred pattern of spread of the Malayo-Polynesian languages, the value is highest in the Philippines, where it amounts to about 28 % of extant lineages, and next in Eastern Indonesia (Sulawesi and the Lesser Sundas), at 19.7 %, falling to 13.6 % in Western Indonesia (Java, Sumatra and Bali) and only 10.3 % in Kalimantan. It is possible that the small groups of dispersing Malayo-Polynesian speakers assimilated indigenous populations in the Philippines before spreading further south and east, although this "demic diffusion" model seems not to apply (despite a longer time frame) within Taiwan.

There are several hints of an increase in population size in the archaeological record after 4 ka that might reflect the expansion of both autochthonous and immigrant groups. First, cave sites with evidence of occupation up to 4 ka, such as Niah Cave in Sarawak and Ulu Leang 1, Leang Burung 1 and Leang Tuwo Mane'e in Sulawesi, also generally show evidence of continued occupation after that date. Second, many cave sites register their initial evidence for occupation between 4 and 3 ka; these include Torongan Cave and Reranum Cave in the Batanes Islands in the Philippines, Arku Cave in Luzon, Bukit Tengkorak in Sabah, Leang Karassak in Sulawesi and Uattamdi in the Moluccas. Third, a new class of sites comprising dated open sites is evident after 4 ka, represented for instance by Dimolit in Luzon and the Kalumpang sites in Sulawesi. There is an increasing number of these sites after 2 ka, for instance the Buni sites in northwest Java, Gilimanuk and Pacung/Sembiran in Bali, and Melolo in Sumba (Bellwood 1997; Bulbeck et al. 2000; Bellwood and Dizon 2014).

Mid-Holocene input into Taiwan from the mainland is also very clear in the founder analysis, and it is especially notable that all of the lineages that showed an "out-of-Taiwan" ancestry in ISEA are also Neolithic markers for the settlement of Taiwan from South China. Although the frequency of these lineages throughout the various aboriginal groups was never greater than 50 % (except in the Ami, with 56 %), it appears therefore that the admixture that generated the current maternal gene structure of Taiwanese tribes took place progressively after the "out-of-Taiwan" expansion. At the time the "out-of-Taiwan" dispersal occurred, the expanding population seems to have remained distinct, with a full South Chinese ancestry (similar to what also seems to have happened to a considerable degree in Central Europe (Haak et al. 2010) and in the Great Lakes region during the Bantu expansion (Gomes et al. 2015; Silva et al. 2015). Whether the subsequent admixture with autochthonous groups (and, presumably, the assimilation of the language) was rapid or protracted we cannot assess from contemporary data alone, in the absence of ancient DNA evidence. Although some hierarchical ancestry for Austronesian languages have been proposed, the most widely accepted Austronesian tree indicates ten basal branches, including Malayo-Polynesian, which on the face of it might support a simultaneous expansion of this hypothetical Austronesian-speaking agriculturalist population across Taiwan and ISEA. However, since the Ami showed the closest similarity to the ancestral Austronesian-speaking population, an expansion across the island from the east coast from the group of Neolithic pioneers who remained on Taiwan might be suggested.

In conclusion, despite the extensive genetic drift, we have shown that a remarkably consistent picture of prehistoric dispersals can be reconstructed from modern mitogenome patterns. By accounting for all of the rarer lineages in ISEA that have a Holocene Taiwanese ancestry, we are now able to provide a definitive, precise description of Holocene maternal ancestry across Taiwan and Southeast Asia. More generally, the success of this procedure illustrates what we believe to be a valuable approach to the study of human dispersals and settlement using mtDNA data: firstly, a founder analysis using large numbers of control-region sequences; followed by the testing of the time depth of individual candidate founder clusters by the more precise means of whole mitogenomes. This procedure is also amenable to use with other non-recombining marker systems, and ultimately, where feasible, to further testing against ancient DNA.

Acknowledgments We thank Freda Oppenheimer, Onchansamone Ninethaphone, Nouphanh Keosouda and Bounheuang Bouasisengpaseuth for the Laos samples, A. S. M. Sofro and John Clegg for Indonesian samples, and Sean O'Riordan for Vietnamese samples. This work was supported by FEDER funds through "Programa Operacional Factores de Competitividade—COMPETE (FCOMP-01-0124-FEDER-029291)" and by national funds through FCT, the Portuguese Foundation for Science and Technology (FCT) through the research project (PTDC/IVC-ANT/4917/2012). AB has a PhD grant from FCT (SFRH/BD/78990/2011). PS is supported by FCT, European Social Fund, Programa Operacional Potencial Humano and the FCT Investigator Programme (IF/01641/2013). IPATIMUP integrates the i3S Research Unit, which is partially supported by FCT. This work is supported by FEDER funds through the Operational Programme for Competitiveness Factors-COMPETE and National Funds through FCT, under the project "PEst-C/SAU/LA0003/2013" (IPATIMUP), byinfrastructure support through NORTE-07-0162-FEDER-00018 and NORTE-07-0162-FEDER-000067, by Programa Operacional Regional do Norte (ON.2—O Novo Norte) through FEDER funds under the Quadro de Referência Estratégico Nacional (QREN) and by FCT I.P. and ERDF (COMPETE 2020- POCI) through the strategic funding programme UID/BIA/04050/2013 (POCI-01-0145-FEDER-007569) (CBMA). MBR thanks the British Academy for financial support through project LRG-42440 and PS and MBR also thank the de Laszlo Foundation and the British Academy (project BARDA-48208).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants All procedures performed in studies involving human participants were approved by the University of Leeds, Faculty of Biological Sciences Ethics Committee, and that of the University of Huddersfield, School of Applied Sciences.

Informed consent Informed consent was obtained from all individuals participants included in the study.

References

Abdulla MA, Ahmed I, Assawamakin A, Bhak J, Brahmachari SK, Calacal GC, Chaurasia A, Chen CH, Chen J, Chen YT, Chu J, Cutiongco-de la Paz EMC, De Ungria MCA, Delfin FC, Edo J, Fuchareon S, Ghang H, Gojobori T, Han J, Ho SF, Hoh BP, Huang W, Inoko H, Jha P, Jinam TA, Jin L, Jung J, Kangwanpong D, Kampuansai J, Kennedy GC, Khurana P, Kim HL, Kim K, Kim S, Kim WY, Kimm K, Kimura R, Koike T, Kulawonganunchai S, Kumar V, Lai PS, Lee JY, Lee S, Liu ET, Majumder PP, Mandapati KK, Marzuki S, Mitchell W, Mukerji M, Naritomi K, Ngamphiw C, Niikawa N, Nishida N, Oh B, Oh S, Ohashi J, Oka A, Ong R, Padilla CD, Palittapongarnpim P, Perdigon HB, Phipps ME, Png E, Sakaki Y, Salvador JM, Sandraling Y, Scaria V, Seielstad M, Sidek MR, Sinha A, Srikummool M, Sudoyo H, Sugano S, Suryadi H, Suzuki Y, Tabbada KA, Tan A, Tokunaga K, Tongsima S, Villamor LP, Wang E, Wang Y, Wang H, Wu JY, Xiao H, Xu S, Yang JO, Shugart YY, Yoo HS, Yuan W, Zhao G, Zilfalil BA (2009) Mapping human genetic diversity in Asia. Science 326:1541–1545. doi:10.1126/science.1177074

Anderson A (2005) Crossing the Luzon Strait: archaeological chronology in the Batanes Islands, Philippines and the regional sequence of Neolithic dispersal. J Austrone Stud 1:25–44

Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147

Bandelt H-J, Forster P, Sykes BC, Richards MB (1995) Mitochondrial portraits of human populations using median networks. Genetics 141:743–753

Barker G, Richards MB (2013) Foraging–farming transitions in Island Southeast Asia. J Archaeol Method Th 20:256–280. doi:10.1007/s10816-012-9150-7

Barker G, Barton H, Bird M, Daly P, Datan I, Dykes A, Farr L, Gilbertson D, Harrisson B, Hunt C, Higham T, Kealhofer L, Krigbaum J, Lewis H, McLaren S, Paz V, Pike A, Piper P, Pyatt B, Rabett R, Reynolds T, Rose J, Rushworth G, Stephens M, Stringer C, Thompson J, Turney C (2007) The 'human revolution' in lowland tropical Southeast Asia: the antiquity and behavior of anatomically modern humans at Niah Cave (Sarawak, Borneo). J Hum Evol 52:243–261. doi:10.1016/j.jhevol.2006.08.011

Bellwood P (1995) Austronesian prehistory in Southeast Asia: homeland, expansion and transformation. In: Bellwood P, Fox JJ, Tryon D (eds) The Austronesians. ANU, Canberra, pp 96–111

Bellwood P (1997) The prehistory of the Indo-Pacific archipelago. University of Hawaii Press, Honolulu

Bellwood P (2005) First farmers. Blackwell, Oxford

Bellwood P, Dizon E (2005) The Batanes Archaeological Project and the "Out of Taiwan" hypothesis for Austronesian dispersal. J Austron Stud 1:1–33

Bellwood P, Dizon E (2008) Austronesian cultural origins. Out of Taiwan, via the Batanes Islands, and onwards to western Polynesia. In: Sanchez-Mazas A, Blench R, Ross M, Peiros P, Lin M (eds) Past human migrations in East Asia matching archeology, linguistics and genetics. Routledge, Great Britain, pp 23–39

Bellwood P, Dizon E (2014) 4000 years of migration and cultural exchange: the archaeology of the Batanes Islands, Northern Philippines. Terra Australis 40, ANU E Press, Canberra

Bellwood P, Fox JJ, Tryon D (2006) The Austronesians: historical and comparative perspectives. ANU E Press, Canberra

Benedict PK (1976) Austro-Thai and Austroasiatic. Ocean Linguist 13:1–36

Bird MI, Taylor D, Hunt C (2005) Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: a savanna corridor in Sundaland? Quat Sci Rev 24:2228–2242. doi:10.1016/j.quascirev.2005.04.004

Bulbeck D (2008) An integrated perspective on the Austronesian diaspora: The switch from cereal agriculture to maritime foraging in the colonisation of Island Southeast Asia. Aust Archaeol 67:31–52

Bulbeck D, Pasqua M, Di Lello A (2000) Culture history of the Toalean of south Sulawesi, Indonesia. Asian Perspect 39:71–108

Chang K-C (1989) The Neolithic Taiwan Strait. Kaogu 6:541–569

Donohue M, Denham T (2010) Farming and language in Island Southeast Asia. Curr Anthropol 51:223–256. doi:10.1086/650991

Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22:1185–1192. doi:10.1093/molbev/msi103

Fagundes NJ, Kanitz R, Eckert R, Valls A, Bogo MR, Salzano FM, Smith DG, Silva WA Jr, Zago MA, Ribeiro-dos-Santos AK (2008) Mitochondrial population genomics supports a single pre-Clovis origin with a coastal route for the peopling of the Americas. Am J Hum Genet 82:583–592

Friedlaender JS, Friedlaender FR, Reed FA, Kidd KK, Kidd JR, Chambers GK, Lea RA, Loo JH, Koki G, Hodgson JA, Merriwether DA, Weber JL (2008) The genetic structure of Pacific Islanders. PLoS Genet 4:e19. doi:10.1371/journal.pgen.0040019

Gomes V, Pala M, Salas A, Alvarez-Iglesias V, Amorim A, Gomez-Carballa A, Carracedo A, Clarke DJ, Hill C, Mormina M, Shaw MA, Dunne DW, Pereira R, Pereira V, Prata MJ, Sanchez-Diz P, Rito T, Soares P, Gusmao L, Richards MB (2015) Mosaic maternal ancestry in the Great Lakes region of East Africa. Hum Genet 134:1013–1027. doi:10.1007/s00439-015-1583-0

Gray RD, Drummond AJ, Greenhill SJ (2009) Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science 323:479–483. doi:10.1126/science.1166858

Haak W, Balanovsky O, Sanchez JJ, Koshel S, Zaporozhchenko V, Adler CJ, Der Sarkissian CS, Brandt G, Schwarz C, Nicklisch N, Dresely V, Fritsch B, Balanovska E, Villems R, Meller H, Alt KW, Cooper A (2010) Ancient DNA from European early Neolithic farmers reveals their Near Eastern affinities. PLoS Biol 8:e1000536. doi:10.1371/journal.pbio.1000536

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: Nucleic Acids Symposium Series, vol 41, pp 95–98

Hill C, Soares P, Mormina M, Macaulay V, Meehan W, Blackburn J, Clarke D, Raja JM, Ismail P, Bulbeck D, Oppenheimer S, Richards M (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol 23:2480–2491. doi:10.1093/molbev/msl124

Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S, Richards M (2007) A mitochondrial stratigraphy for Island Southeast Asia. Am J Hum Genet 80:29–43. doi:10.1086/510412

Jinam TA, Hong LC, Phipps ME, Stoneking M, Ameen M, Edo J, Saitou N (2012) Evolutionary history of continental Southeast Asians: "Early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. Mol Biol Evol. doi:10.1093/molbev/mss169

Karafet TM, Hallmark B, Cox MP, Sudoyo H, Downey S, Lansing JS, Hammer MF (2010) Major east-west division underlies Y chromosome stratification across Indonesia. Mol Biol Evol 27:1833–1844. doi:10.1093/molbev/msq063

Kayser M, Choi Y, van Oven M, Mona S, Brauer S, Trent RJ, Suarkia D, Schiefenhovel W, Stoneking M (2008) The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia. Mol Biol Evol 25:1362–1374. doi:10.1093/molbev/msn078

Ko AM-S, Chen C-Y, Fu Q, Delfin F, Li M, Chiu H-L, Stoneking M, Ko Y-C (2014) Early Austronesians: into and out of Taiwan. Am J Hum Genet 94:426–436

Loo J-H, Trejaut JA, Yen J-C, Chen Z-S, Lee C-L, Lin M (2011) Genetic affinities between the Yami tribe people of Orchid Island and the Philippine Islanders of the Batanes archipelago. BMC Genet 12:21

Macaulay V, Hill C, Achilli A, Rengo C, Clarke D, Meehan W, Blackburn J, Semino O, Scozzari R, Cruciani F, Taha A, Shaari NK, Raja JM, Ismail P, Zainuddin Z, Goodwin W, Bulbeck D, Bandelt HJ, Oppenheimer S, Torroni A, Richards M (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308:1034–1036. doi:10.1126/science.1109792

O'Connor S, Bulbeck D (2014) Homo sapiens societies in Indonesia and South-Eastern Asia. The Oxford Handbook of the Archaeology and Anthropology of Hunter-Gatherers. Oxford University Press, Oxford, pp 346–367

Olsen JW, Miller-Antonio S (1992) The Palaeolithic in southern China. Asian Perspect 31:129–160

Oppenheimer S (1998) Eden in the east: the drowned continent of Southeast Asia. Phoenix, London

Pereira L, Freitas F, Fernandes V, Pereira JB, Costa MD, Costa S, Maximo V, Macaulay V, Rocha R, Samuels DC (2009) The diversity present in 5140 human mitochondrial genomes. Am J Hum Genet 84:628–640. doi:10.1016/j.ajhg.2009.04.013

Pereira L, Silva NM, Franco-Duarte R, Fernandes V, Pereira JB, Costa MD, Martins H, Soares P, Behar DM, Richards MB, Macaulay V (2010) Population expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. BMC Evol Biol 10:390

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251–1276

Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, Soares P (2013) The first modern human dispersals across Africa. PLoS One 8:e80031

Ross M (2005) The Batanic languages in relation to the early history of the Malayo-Polynesian subgroup of Austronesian. J Austron Stud 1:1–24

Saillard J, Forster P, Lynnerup N, Bandelt H-J, Nørby S (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. Am J Hum Genet 67:718–726

Silva M, Alshamali F, Silva P, Carrilho C, Mandlate F, Jesus Trovoada M, Cerny V, Pereira L, Soares P (2015) 60,000 years of interactions between Central and Eastern Africa documented by major African mitochondrial haplogroup L2. Sci Rep 5:12526. doi:10.1038/srep12526

Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, Macaulay V, Bulbeck D, Oppenheimer S, Lin M, Richards MB (2008) Climate change and postglacial human dispersals in Southeast Asia. Mol Biol Evol 25:1209–1218. doi:10.1093/molbev/msn068

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740–759

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke DJ, Loo J-H, Thomson N, Denham T, Donohue M, Macaulay V, Lin M, Oppenheimer S, Richards MB (2011) Ancient voyaging and Polynesian origins. Am J Hum Genet 88:239–247. doi:10.1016/j.ajhg.2011.01.009

Soares P, Alshamali F, Pereira JB, Fernandes V, Silva NM, Afonso C, Costa MD, Musilova E, Macaulay V, Richards MB, Cerny V, Pereira L (2012) The expansion of mtDNA haplogroup L3 within and out of Africa. Mol Biol Evol 29:915–927. doi:10.1093/molbev/msr245

Soares P, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, Mormina M, Brandão A, Fraser RM, Wang T-Y, Loo J-H, Snell C, Ko T-M, Amorim A, Pala M, Macaulay V, Bulbeck D, Wilson JF, Gusmão L, Pereira L, Oppenheimer S, Lin M, Richards MB (2016) Resolving the ancestry of Austronesian-speaking populations. Hum Genet. doi:10.1007/s00439-015-1620-z

Solheim WG (2006) Archaeology and culture in Southeast Asia: unraveling the Nusantao. University of Philippines Press, Quezon City (Philippines)

Tabbada KA, Trejaut J, Loo JH, Chen YM, Lin M, Mirazon-Lahr M, Kivisild T, De Ungria MC (2010) Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? Mol Biol Evol 27:21–31. doi:10.1093/molbev/msp215

Trejaut JA, Kivisild T, Loo JH, Lee CL, He CL, Hsu CJ, Lee ZY, Lin M (2005) Traces of archaic mitochondrial lineages persist

in Austronesian-speaking Formosan populations. PLoS Biol 3:e247. doi:10.1371/journal.pbio.0030247

Trejaut JA, Poloni ES, Yen J-C, Lai Y-H, Loo J-H, Lee C-L, He C-L, Lin M (2014) Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia. BMC Genet 15:77

Tumonggor MK, Karafet TM, Hallmark B, Lansing JS, Sudoyo H, Hammer MF, Cox MP (2013) The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific. J Hum Genet 58:165–173

van Oven M, Kayser M (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat 30:E386–E394

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS 13:555–556

STUDY OF AUTOSOMAL VARIATION IN SOUTHEAST ASIA

Paper III

The genome-wide landscape of Island Southeast Asia

PAPER III

The genome-wide landscape of Island Southeast Asia

Brandão A, Cavadas B, Bulbeck D, Hudson B, White J, Chia S, Saidin M, Zafarina Z, Oppenheimer S, Pereira L, Richards MB, Soares P. The genome-wide landscape of Island Southeast Asia. *In preparation*

The genome-wide landscape of Island Southeast Asia

Brandão A[1,2,3,4],, Cavadas B[1,2], Bulbeck D[5], Hudson B[6], White J[7], Chia S[8], Saidin M[8], Zafarina Z[9,10], Oppenheimer S[11], Pereira L[1,2,12], Richards MB[3,13], Soares P [1,2,13,14]

[1]IPATIMUP (Institute of Molecular Pathology and Immunology of the University of Porto), Rua Júlio Amaral de Carvalho,45, 4200-135 Porto, Portugal

[2]i3S (Instituto de Investigação e Inovação em Saúde, Universidade do Porto), 4200 Porto, Portugal.

[3]Department of Biological Sciences, School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, United Kingdom

[4]ICBAS (Instituto Ciências Biomédicas Abel Salazar), Universidade do Porto, Rua de Jorge Viterbo Ferreira n.º 228, 4050-313 Porto, Portugal.

[5]Department of Archaeology and Natural History, College of Asia and the Pacific, The Australian National University, Acton ACT 2601, Canberra, Australia

[6]Archaeology Department, University of Sydney, New South Wales 2006, Australia

[7]Dept. of Anthropology, University of Pennsylvania Museum, 3260 South St. Philadelphia, United States of America.

[8]Centre for Archaeological Research, Universiti Sains Malaysia, 11800 USM Penang, Malaysia

[9]Malaysian Institute of Pharmaceuticals and Nutraceuticals Malaysia, National Institutes of Biotechnology Malaysia, Penang, Malaysia

[10]Human Identification Unit, School of Health Sciences, Health Campus, University Sains Malaysia, Penang, Malaysia

[11]Institute of Human Sciences, School of Anthropology, University of Oxford, The Pauling Centre, 58a Banbury Road, Oxford OX2 6QS, United Kingdom

[12]Faculty of Medicine, University of Porto, Al. Prof. Hernâni Monteiro, 4200 - 319 Porto, Portugal

[13]Faculty of Biological Sciences, University of Leeds, LS2 9JT Leeds, United Kingdom

[14]CBMA (Centre of Molecular and Environmental Biology), Department of Biology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

Report

The demographic events that shaped the complex human genetic architecture of Island Southeast Asia (ISEA) are still surrounded by controversy. The extensive ethno-linguistic and genetic variation of ISEA populations has likely been significantly established by multiple migrations and population expansions throughout the history of the region. The most accepted model for decades has been the “Out-of-Taiwan”, based mostly on linguistic evidence and additionally supported to some extent by archaeological data[1; 2]. The model stated that following the first settlement by the ancestors of the so-called “Australoid” populations (found nowadays scattered around SEA and Australia) at least 50 thousand years ago (ka), a massive agriculture-driven migration in the mid-Holocene 4 ka ago displaced previous settlers and constitute the major ancestry of present day ISEA populations[1; 2]. The model refers to a putative dispersal of Neolithic farmers from South China to Taiwan around 6 ka and onward to ISEA and the Pacific, carrying a new cultural package and the Austronesian languages that are spoken in the region nowadays[1-4].

Recent genetic studies suggested that the two-wave hypothesis is too simplistic and does not account for the complex genetic variation observed in Southeast Asian (SEA) populations. In fact, climatic changes have been appointed by many as the major driving force for the population movements in the region[5-10]. SEA is currently divided into Mainland Southeast Asia (MSEA) and ISEA but until the Last Glacial Maximum (LGM) approximately 20 ka, it was partially joined together by an extensive landmass known as Sundaland linking Asian mainland and the current islands of Sumatra, Java and Borneo[11; 12], mirroring the Sahul continent that split into present-day Australia and New Guinea. Following the LGM, climate warming led to three major episodes of sea level rise (at ~14.5, 11.5, and probably ~7.5 ka), which caused the flooding of almost half of the land area of the Sundaland continent and doubled the length of the coastline[11; 12]. These dramatic changes likely trigged drastic population displacements in the region[5-10; 13-16]. Such events would predate but do not exclude an OOT event. Recently, using both mtDNA and Y-chromosome, as well as a reanalysis of published genome-wide data, we proposed a model where the postglacial expansions possibly played the major role in establishing the population genetic structure of the region, but a clear low-scale OOT migration was detected and it could have been responsible for the spread of the Austronesian languages and even some technological innovations[8; 17].

In the last two decades most genetic studies on ISEA relied mainly on the variation in the haploid markers mtDNA and non-recombining Y-chromosome[5-7; 9; 13; 15]. More recently, the development of high-throughput single-nucleotide polymorphism (SNP)

genotyping technologies allowed for more comprehensive genome-wide studies to be employed[18-21], many of them focused on Malaysia, however they are overall still scarce. The most comprehensive study was performed by the Pan-Asian SNP Consortium[18] that corroborated that all South, East and Southeast Asian populations had their genesis on a single migration 'Out-of-Africa' via a southern coastal route[22] and revealed genetic stratification among East and Southeast Asians populations[18]. Genome-wide studies addressing the complex genetic architecture of SEA populations, have supported both OOT-type models[23; 24] and multiple-wave[21; 25] migration models. However, caution is needed when interpreting these results due to limitations in sampling (some lack of significantly representation of some populations in SEA and Taiwan) or the interpretative limitations of some analyses.

In order to provide a comprehensive genetic ancestry landscape of SEA populations, without some of the drawbacks of previous studies namely on the low number of SNPs analysed[18; 24] or limitation in sampling[19; 20], we conducted an in-depth population genetic study using genome-wide SNP data from 47 populations (including newly genotyped together with samples from the 1000 Genomes Project[26] and HGDP-CEPH repository[27]) spread throughout SEA and neighbouring East Asia and Pacific regions. By exploring the genome-wide diversity in a large number of populations, we attempt to contribute new insights on the genetic relationships between the populations existing in the region and shed light on the past events that shaped the genetic makeup of the region and the massive ethno-linguistic diversity seen in present-day Southeast Asia.

We genotyped 416 individuals using the Illumina HumanOmniExpress BeadChip, containing ~700,000 SNPs. We excluded 48 individuals that exhibited distinct admixture composition compared with samples of the same populations (outliers). In addition, we also removed possible duplicate samples and first-degree relatives (63 individuals). In the SNPs filtering we excluded markers with genotyping call rates <95% and minor allele frequency <5%, using PLINK software[28]. A total of 305 individuals, representing 34 populations from Southeast and the Pacific, and 572,502 SNPs passed the quality control. In order to increase spatial resolution, we merged our primary dataset with data from relevant populations from the 1000 Genomes Project26 and from the HGDP-CEPH repository[27], obtaining a final dataset of 644 individuals from 47 populations typed at 258,994 overlapping SNPs (details on the analysed populations can be found in Table S1). Several analyses require a minimum background linkage disequilibrium (LD) so we obtained a pruned dataset by removing SNPs with excessive LD (pairwise genotypic correlation $r2 > 0.4$), within a 50-SNP sliding window with step of 10 SNPs. The resulting pruned dataset contained 139,737 SNPs. Many of the published genome-wide results

were based on the Pan-Asian SNP Consortium dataset[8; 18; 24; 25], so we will compare the obtained results with a reanalysis of this data. We added the relevant 1000 Genomes Project populations[26] resulting in an expanded Pan-Asian SNP Consortium (EPASC) dataset with 19,338 SNPs after pruning (Table S2).

Principal component analysis (PCA) was used to reduce the data into vectors that capture the main trends across East/Southeast Asian populations. PCA was carried out in the pruned dataset by using the smartpca tool, included in the EIGENSOFT v6.0.1 package29. We employed ADMIXTURE v1.23 to estimate population structure using maximum likelihood (ML)[30], with ten-fold cross validation to find the optimal number of clusters or ancestral populations (K). For comparison we used sNMF v1.2 which estimates genetics components using a model-free approach based on a non-negative matrix factorization[31]. PLINK v1.928 was used to estimate Runs of Homozygosity (ROH) and the Inbreeding Coefficient (IC) in the unpruned dataset of selected Southeast Asian populations. The ROH was calculated taking 5,000 kb (50 SNPs) sliding windows across the genome and allowing for one heterozygous and five missing calls in each window. Two consecutive ROHs are considered as a single unit if their distance is ≤ 1 Mb. The TreeMix v1.1226[32] was used to infer relationships and migration events between selected populations. The estimated ML tree was rooted by an Indian population, and blocks of 500 SNPs (-k 500) to account for LD and migration edges were added sequentially until the model explained ~99.7 % of variance.

The overall results of the PCA analysis are shown in Figure 1. To simplify the spatial disposition of the populations, they were grouped and labelled according to their general geographic location (the PCA displaying all individualized populations is found in Figure S1). The first component (PC1), capturing 27.03 % of total variation, clusters the Southeast Asian and New Guinean populations in a west-east axis with East Indonesian populations (Alor, Palu and Ambon) in an intermediate position. The second component (PC2) at 6.43 % shows a geographic clustering of populations between North of East Asia on one extreme and SEA on the other. These two components, north-south and west-east, provide a general genetic clustering that closely resembles the geography as displayed in figure 1. Interestingly, from both first and second components, Myanmar (formerly Burma) appear to be differentiated from the other MSEA populations on PC1 which can be caused by some level of South Asian ancestry, and it is more closely related to China on PC2 which is not fully unexpected on geographical terms. Taiwanese aboriginals are probably the populations showing the major discrepancy between the plot and the geography. Although located offshore of South China, they occupy a position close to the extreme of the PC2 variation.

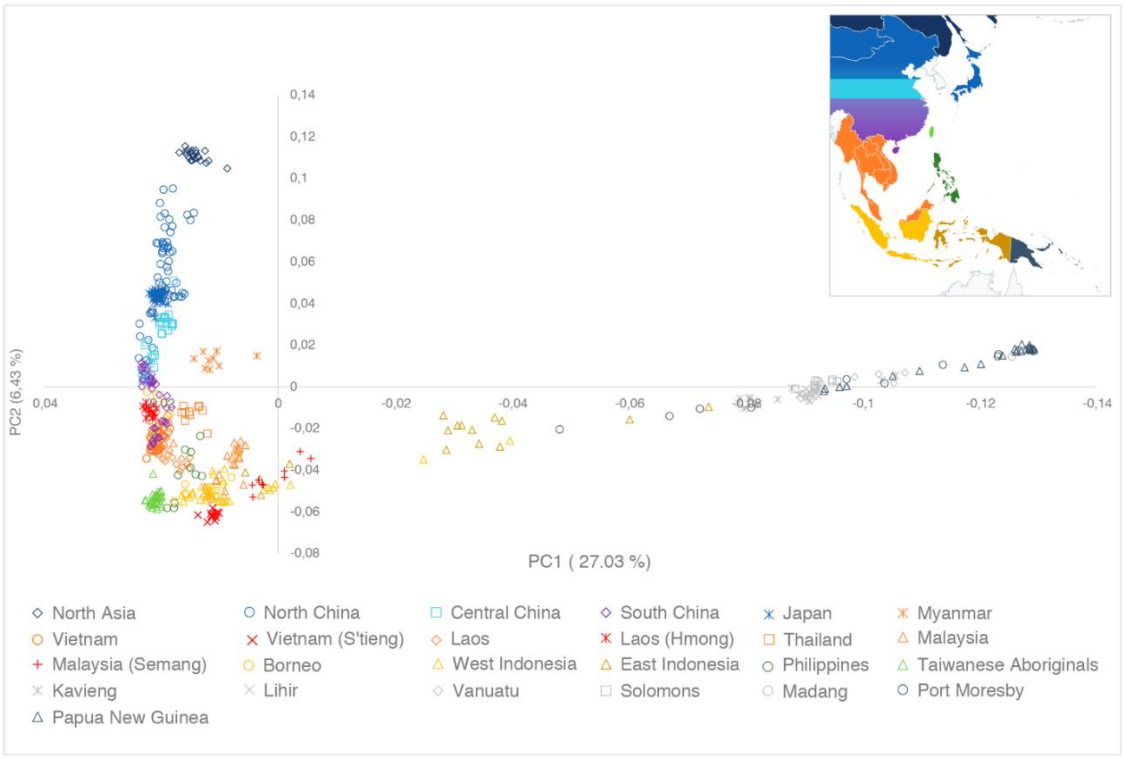


Figure 1. Principal Component Analysis of Southeast Asian populations and surrounding Asian populations. The populations are grouped according their geography, in which North Asia comprises the Yakut, North China comprises Mongolia, Hezhen, Daur and Oroqen; Central China comprises Naxi, Tujia, Yizu, Tu and Central Han; South China comprises Dai; Miao; Lahu, She and Southern Han; Taiwanese Aboriginals comprises Atayal, Bunun, Ami and Paiwan; East Indonesia comprises Alor, Palu and Ambom; West Indonesia comprises the populations from Borneo island (Kota Kinabalu, Banjarmasin and Palangkaraya) and Bali, Mataram and Sumatra (Palembang, and Pekanbaru). PCA analysis was paired with the geographic map of the samples.

To evaluate the ancestral populations or components that may have shaped the genetic structure in present SEA populations, we applied ADMIXTURE analysis (Figure 2). sNMF was also used as a cross-checking of the estimated genetic structure (Figure S2). Figure 2 shows the ADMIXTURE results from K=3 to K=8 (with K=7 identified as the "optimal" number of clusters in the dataset by cross-validation; Figure S3). For K=3 the three components separate East Asian Populations (blue), Oceanic populations (grey), and the African populations (green). South Asian populations (green) appear as a mixture of African and East Asian. At K=4 a South Asian component emerges and at K=5 the East Asian component differentiates into a northerner component (blue) and a southerner component (orange) with progressive changes in frequency from the Yakut in Siberia to SEA. At K=6 the Oceanic component splits into a New Guinean component and a component corresponding to the offshore islands of New Guinea (brown) although this split does not exist in further analyses for higher values of K. Given this, at K=7, the

putative optimal K from cross-validation, the predominantly SEA component at K=5 and K=6 becomes three components: the orange one is widely frequent from Japan to MSEA displaying frequencies of about 50% through a wide range of these populations with nearly 100% in the Hmong; the yellow component mainly relates to MSEA ethnic groups, namely the Malaysian Semang and Vietnamese Stieng, but it is present at discernible frequencies at most MSEA populations; and a dark blue component mostly frequent in Taiwan and ISEA populations, but also with minor frequencies in Vietnam, Laos and South China. We also estimated ancestry using sNMF from K=3 to 7 (Figure S2) and although the principles behind the algorithm are very different from ADMIXTURE, the results strongly match the ones in ADMIXTURE providing confidence on the ancestry estimates.

Above the hypothetical optimal K, at K=8, a new component (light blue) appears, differentiating the Hmong population from the other Laos populations. The high values of this population in term of ROH and IC suggests a high degree of population isolation from the neighbouring populations (Figure S4). Interestingly in the EPASC dataset at higher values of K, Hmong also develop a specific component that is present in Hmong-speaking populations (data not shown) independently of their geography suggesting a common and recent ancestry for populations of this linguistic group. A further look at some details in the ADMIXTURE analysis reveals some residual Indo-European component (dark green) in Malaysia (14.6%) and Myanmar (21.3%), whereas a MSEA component is found in India (20.9%), suggesting some bidirectional gene flow between MSEA and Indian populations. Traces of population interaction between Indian and SEA populations, within the Bay of Bengal, have been suggested before, and it could be associated with dispersal of Austroasiatic languages[33]. Considering K=7, most of MSEA populations (Vietnam, Thailand, Laos and Myanmar) with the exception of the Malays have about 50% of the widespread East Asian/Southeast Asian component suggesting a shared genetic ancestry of what is basically half of the genetic variation of East Asia. This component is hardly present in ISEA. The S'tieng in Vietnam and the Semang in Malaysia share a component (yellow) at nearly 100% that is mostly restricted to MSEA. While the predominance of this component in these two populations could be a signal of high drift, both the ROH and the IC display values that do not support such a drastic reduction of diversity when compared with other populations (Figure S4). One alternative is that these populations, commonly included in the so-called relict groups of SEA might reflect a deeper layer of SEA genetic diversity. The Semang and specially the Stieng have also an extreme position in the PC2 (corresponding to the North-SEA axis) suggesting their status as most genuine SEA populations eventually with ISEA populations (as we will discuss below). The other MSEA

populations show an intermediate position in this PC2 between these groups and China (or South China) probably suggesting a dual ancestry. This is reflected on the ADMIXTURE and sNMF ancestry estimates, where MSEA populations share a mixed ancestry corresponding to an ancestral component (the yellow and/or dark blue components) and an incursion of a component (orange) of possible Chinese origin. This orange component might suggest a putative South Chinese expansion either motivated by climate or Neolithic practices. The calculation of admixture times using ALDER[34] always led to dates in the last few centuries, probably extremely underestimated, and we opted to exclude them.

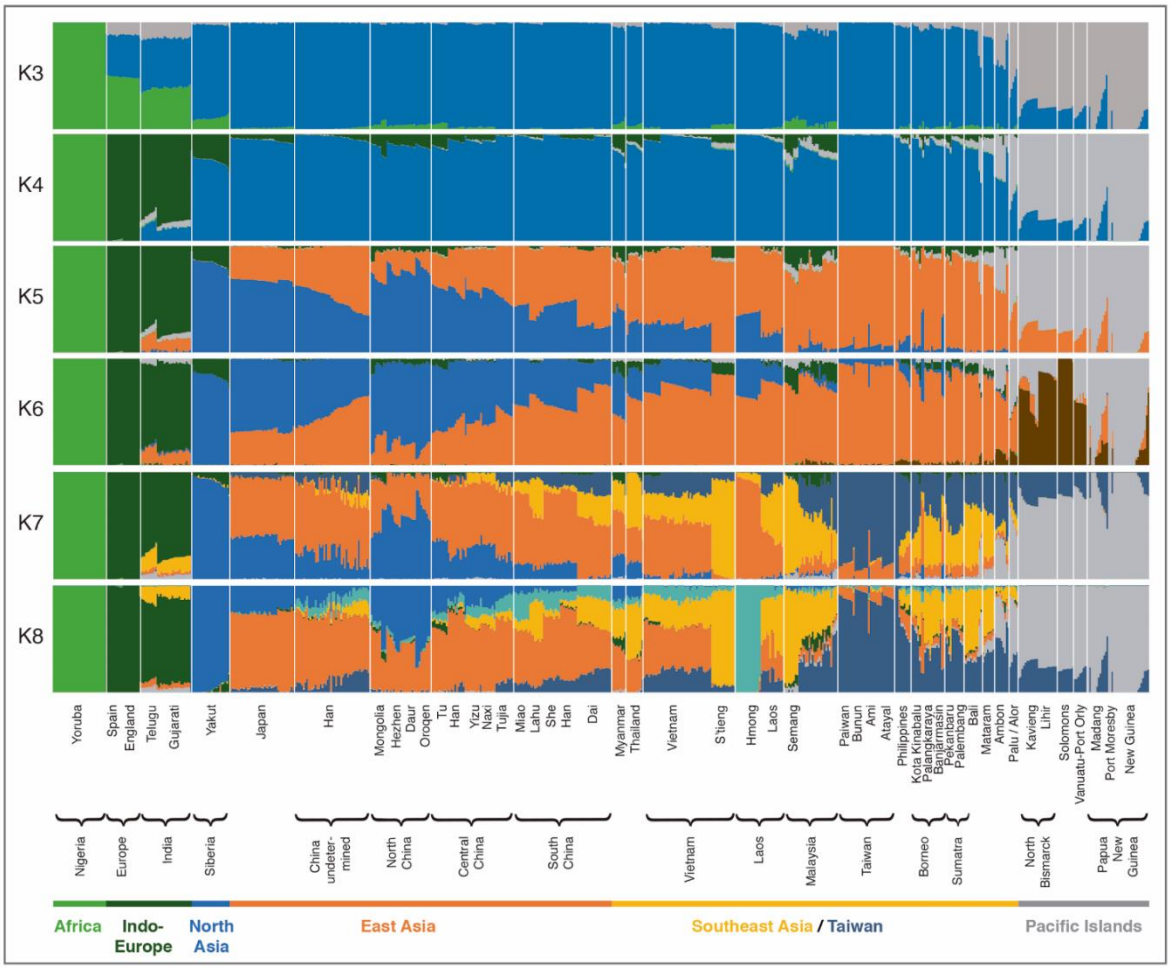


Figure 2. Admixture analysis (K=2 to K=8) of Asian and neighbouring populations and subpopulations. Each individual from populations is represent in the x-axis, as a vertical stacked column of color-coded admixture proportions of the putative ancestral populations.

ADMIXTURE and sNMF ancestry analyses revealed that the region of ISEA harbours a very specific genetic structure. There is a clear genetic cline between West and East ISEA with the first one showing components representative of MSEA while East

Indonesia clearly shows an input from Papuan-speaking populations. These two genetic influxes are already attested from archaeology, linguistic and genetics[8; 13; 14; 24; 35-38] although with a few variations on their timing. This leaves the dark blue component as the most representative ISEA ancestry. This component is more frequent in Taiwan (about 90%) but it is also frequent throughout all of ISEA and it is also present at low frequencies in the Pacific and on MSEA. It is mostly present in Austronesian-speaking populations which could render it an OOT-ancestry marker. However higher frequencies in Taiwan do not indicate source, and no directionality of the spread of the component is visible. If the major component of genetic variation in ISEA was established at least in the early Holocene as suggested by uniparental markers[5; 6; 8; 9; 14], and it was followed by gene flow from the west (MSEA) and the east (Papuan), the frequency of the Holocene component would decrease, which renders any inference based on frequencies inconclusive. In contrast, in Taiwan we just observe a minor genetic input of the mainland Orange component.

We can draw some inferences from the already described analyses regarding the ancestry of the dark blue component. In terms of PCA we can observe the positioning of Taiwan in the North-South axis that mostly relates Asian populations. Considering an OOT model we would expect a large-scale migration of rice-agriculturalists from South China into Taiwan about 6-7 ka ago following a migration from further North where rice domestication had its probable origin[2]. In this model we would expect Taiwan to show some level of similarity with South China but in the PCA they appear on the extreme of the variation of the PC2, close to relict groups like the Stieng and the Semang, suggesting a deeper differentiation from mainland Asia in disagreement with the geographically proximal China and the “out-of-Taiwan” model. MSEA populations follow the cline from North Asia and in terms of structure they display a high frequency of the orange component that is highly frequent in South China. In the EPASC dataset, Oceania is hardly represented but PC1 separates populations from North China to the population labelled Melanesia (Figure S5). PC2 establishes a wide range of diversity with the Melanesian population at one extreme and the Malaysian Negrito group on the other. Nevertheless both main PCs establish a cline in continental populations from Japan/Korea to MSEA while ISEA and Peninsular Malaysian populations appear further differentiated from this more homogenous group. In this case Taiwan appears slightly closer to this group but still differentiated which makes sense as that Taiwanese sample set showed a higher frequency of a South Chinese component (25-30%) than what we are obtaining here for the orange component (10-15%). This can be caused by bias in the SNPs used in each analysis. Still neither of the ancestry analyses performed indicate directionality. The

other hypothesis is that as suggested by haploid markers this could represent a postglacial expanding component that reached Taiwan. Although the component is also present in MSEA and South China it is mostly present on Eastern Coastal populations suggesting that it could be related to major maritime expansions. On mtDNA many of the haplogroups that expanded in the postglacial in ISEA were related with continental China and MSEA at a deeper depth characterized by longer branches between the split with continental Asia and the expanding clade[17]. This can explain the positioning of ISEA and Taiwanese groups in the PCA within the continuum of the North-South axis but further apart from the continent.

We used TreeMix, that allows establishing evolutionary relationships between populations and, importantly, allows inferring admixture between populations by establishing migrations between groups. Six migration edges were added until the model explained 99.7 % of the variance (Figure S6). Using an Indian population to root the ML tree, the populations separated into two groups, one consisting of Pacific populations and the other consisting on Asian populations (Figure 3). Following the results obtained by the previous ancestry estimates (PCA, ADMIXTURE and sNMF) the presence of the Pacific component in East Indonesia is clear as it is shown by a migration with a high weight. The remaining part of the tree corresponds to East/Southeast Asia. There a first basal split between the Malay and the remaining populations, which again suggests some level of antiquity of its components, the yellow and dark blue in the ancestry analyses. Then the tree splits into two groups, one containing SEA populations (both MSEA and ISEA), again groups mostly represented by these same two components. On a note, Laos and Thailand that are present in this cluster contain a high frequency of the orange component but migration with high weight is visible from the East Asian group to this. Another split is obtained between a group containing Austronesian-speaking populations (including the Taiwanese groups) mostly represented by the dark blue component in ADMIXTURE and a group that is mainly East Asian with some MSEA populations, mainly represented by the orange and blue components, which again can mimic the split observed in mtDNA between most haplogroups present in ISEA and the mainland in the late Pleistocene/early Holocene[17]. Overall the tree suggests a deeper ancestry of SEA populations (both MSEA and ISEA) in relation to East Asian populations, where East Asian diversity is gathered in a single clade with the SEA diversity present at deeper clades. This would be expected following a first colonization of SEA in the context of a Southern Coastal Route model for the Out of Africa migration[18; 22; 39] followed by a settlement of East Asia from SEA.

The analysis of the tree in the context of considering an “Out-of-Taiwan” or mostly a Postglacial expansion needs to be carefully considered in two ways, in terms of tree

structure and attending to the migration estimates. In terms of tree structure Taiwan aboriginals are embedded as a single clade within a cluster of ISEA and Austronesian-speaking populations. This strongly suggests that the diversity in Taiwan is a subset of the ISEA diversity as it was obtained previously in a tree that did not consider admixture and migration[18]. Although we grouped the four Taiwanese groups in two groups, the results are similar when using the four separately and they constitute a single sub-clade. One could argue that Taiwanese tribes went through high genetic drift but considering that their history is of independent drift due to isolation from each other there would be no reason that they would be clustered in the same clade unless they correspond to the same subset of variation.

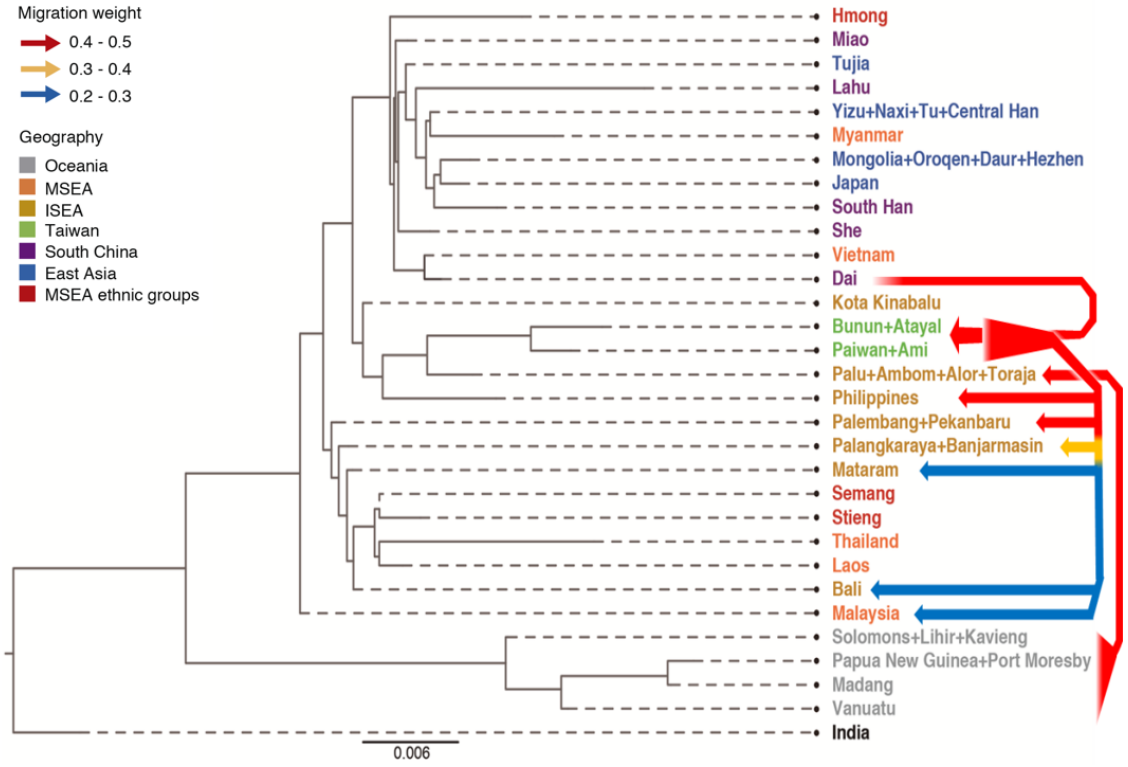


Figure 3. Maximum likelihood population tree and admixture events inferred by TreeMix. The tree that best fit the dataset has six inferred migration edges, which explains 99.7% of genetic variation of the populations. The tree displays the relevant detected migrations involving ISEA and Taiwan. For the full migrations' results involving all the dataset see figure S6. The spectrum of colour of the migration arrows indicates different migration weights. The branch lengths are proportional to the amount of genetic drift that has occurred on populations.

On the other hand high weight migrations were detected from Taiwan into the Philippines and Sumatra but also lower weight migrations into other parts of ISEA. The most probable interpretation for this pattern, one that is fully supported by our recent analysis of haploid markers[8; 17], is that the position of Taiwanese groups in the tree is reflecting a general ancestry in ISEA that according to mtDNA could have been established by postglacial expansions in the early to mid-Holocene that reached Taiwan in the process[6; 8]. However the migration pattern estimated shows that a migration from Taiwan into ISEA occurred against the general pattern that could match an "Out-of-Taiwan" migration. This pattern fits perfectly our most recently established model that suggests a more ancient single genetic ancestry of ISEA and Taiwan followed by a much lower scale Out-of-Taiwan migration for the spread of Austronesian languages, that was mostly representative in the Philippines and that could have spread across ISEA mostly based on the spread of elite groups and language shift of the major ISEA population, rather than large-scale population replacement[38; 40].

The two higher weight migrations from Taiwan into ISEA correspond to the Philippines and Sumatra where we detected higher frequencies of mtDNA haplogroup M7c3c, the most representative putative clade for the Out-of-Taiwan in mtDNA[8]. Our results oppose a previous analysis, based in part on the HUGO-Pan Asia SNP dataset[24] that suggested an ancestry of Austronesian populations in Taiwan. We used our EPASC dataset to estimate a tree with admixture with Treemix (Figure S7) and the two Taiwanese groups again appear as a single minor clade deeply engrained within a broader Austronesian clade. On this dataset no Out-of-Taiwan migration was detected but the resolution in terms of SNPs is substantially lower. Also it displays the clear high diversity of SEA populations in relation to East Asian, including the so-called Philippine Negritos not available in our dataset.

Considering the mtDNA picture[8; 17], this mid-Holocene migration followed a migration from continental China whose population remained relatively isolated until the migration into ISEA. In the genome-wide data it is difficult to access this pattern. A migration from continental Asia was detected with high weight from the Dai population in South China that can be responsible for the presence of the orange component in Taiwan and even further South in the Philippines and the remaining ISEA which can represent a similar pattern to mtDNA[17], but the presence of this component is found at lower frequencies than the ones estimated from mtDNA and Y-chromosome[8; 17]. The putative mtDNA in general seem to show a somewhat Northern Chinese origin[17] although we cannot trace such pattern in our dataset. Still, M7c3c shows an immediate origin in South

China[8] and the major putative “Out-of-Taiwan” Y-chromosome clade in ISEA, O1a2[8] within haplogroup O1a might have had an origin in South China[7; 41].

Overall the results display a complex pattern that cannot be explained by simplistic two-tier models that prevailed for a long time in the field. The results also reveal SEA, including ISEA, as a very diverse region and a reservoir of ancient Asian genetic diversity. As previously suggested, the great ancestry of ISEA dates to substantially earlier than the mid-Holocene although a “Out-of-Taiwan” migration might have been responsible for the introduction of Austronesian languages whose expansion in ISEA might have taken place through language shift of the indigenous population[8; 17; 40].

Reference

1. Bellwood, P. (1997). Prehistory of the Indo-Malaysian archipelago.(ANU E Press).
2. Bellwood, P. (2004). First Farmers: The Origins of Agricultural Societies.(Wiley).
3. Blust, R. (1976). Austronesian culture history: some linguistic inferences and their relations to the archaeological record. World Archaeology 8, 19-43.
4. Blust, R. (1995). The prehistory of the Austronesian-speaking peoples: a view from language. Journal of World Prehistory 9, 453-510.
5. Capelli, C., Wilson, J.F., Richards, M., Stumpf, M.P.H., Gratrix, F., Oppenheimer, S., Underhill, P., Pascali, V.L., Ko, T.M., and Goldstein, D.B. (2001). A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. American Journal of Human Genetics 68, 432-443.
6. Soares, P., Trejaut, J.A., Loo, J.H., Hill, C., Mormina, M., Lee, C.L., Chen, Y.M., Hudjashov, G., Forster, P., MacAulay, V., et al. (2008). Climate change and postglacial human dispersals in Southeast Asia. Molecular Biology and Evolution 25, 1209-1218.
7. Karafet, T.M., Hallmark, B., Cox, M.P., Sudoyo, H., Downey, S., Lansing, J.S., and Hammer, M.F. (2010). Major east-west division underlies Y chromosome stratification across Indonesia. Molecular Biology and Evolution 27, 1833-1844.
8. Soares, P., Trejaut, J.A., Rito, T., Cavadas, B., Hill, C., Eng, K.K., Mormina, M., Brandão, A., Fraser, R.M., Wang, T.-Y., et al. (2016). Resolving the ancestry of Austronesian-speaking populations. Human Genetics. 35, 309-326.
9. Hill, C., Soares, P., Mormina, M., Macaulay, V., Clarke, D., Blumbach, P.B., Vizuete-Forster, M., Forster, P., Bulbeck, D., Oppenheimer, S., et al. (2007). A mitochondrial stratigraphy for Island Southeast Asia. American Journal of Human Genetics 80, 29-43.
10. Oppenheimer, S. (1998). Eden in the east: the drowned continent of Southeast Asia.(Phoenix London).
11. Pelejero, C., Kienast, M., Wang, L., and Grimalt, J.O. (1999). The flooding of Sundaland during the last deglaciation: Imprints in hemipelagic sediments from the southern South China Sea. Earth and Planetary Science Letters 171, 661-671.
12. Blanchon, P., and Shaw, J. (1995). Reef drowning during the last deglaciation: evidence for catastrophic sea-level rise and ice-sheet collapse. Geology 23, 4-8.
13. Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., Clarke, D., Raja, J.M., Ismail, P., Bulbeck, D., et al. (2006). Phylogeography and

ethnogenesis of aboriginal Southeast Asians. Molecular Biology and Evolution 23, 2480-2491.

14. Soares, P., Rito, T., Trejaut, J., Mormina, M., Hill, C., Tinkler-Hundal, E., Braid, M., Clarke, D.J., Loo, J.H., Thomson, N., et al. (2011). Ancient voyaging and polynesian origins. American Journal of Human Genetics 88, 239-247.
15. Tumonggor, M.K., Karafet, T.M., Hallmark, B., Lansing, J.S., Sudoyo, H., Hammer, M.F., and Cox, M.P. (2013). The Indonesian archipelago: An ancient genetic highway linking Asia and the Pacific. Journal of Human Genetics 58, 165-173.
16. Solheim, W.G. (2006). Archaeology and culture in Southeast Asia: unraveling the Nusantao.(UP Press).
17. Brandão, A., Eng, K.K., Rito, T., Cavadas, B., Bulbeck, D., Gandini, F., Pala, M., Mormina, M., Hudson, B., White, J., et al. (2016). Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia. Human Genetics (in press). DOI: DOI 10.1007/s00439-016-1640-3.
18. Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S.K., Calacal, G.C., Chaurasia, A., Chen, C.H., Chen, J., Chen, Y.T., et al. (2009). Mapping human genetic diversity in Asia. Science 326, 1541-1545.
19. Hatin, W., Nur-Shafawati, A., Etemad, A., Jin, W., Qin, P., Xu, S., Jin, L., Tan, S.-G., Limprasert, P., Feisal, M., et al. (2014). A genome wide pattern of population structure and admixture in peninsular Malaysia Malays. The HUGO Journal 8, 5.
20. Aghakhanian, F., Yunus, Y., Naidu, R., Jinam, T., Manica, A., Hoh, B.P., and Phipps, M.E. (2015). Unravelling the Genetic History of Negritos and Indigenous Populations of Southeast Asia. Genome Biology and Evolution 7, 1206-1215.
21. Deng, L., Hoh, B., Lu, D., Fu, R., Phipps, M., Li, S., Nur-Shafawati, A., Hatin, W., Ismail, E., Mokhtar, S., et al. (2014). The population genomic landscape of human genetic structure, admixture history and local adaptation in Peninsular Malaysia. Human Genetics 133, 1169-1185.
22. Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., et al. (2005). Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science 308, 1034-1036.
23. Xu, S., Pugach, I., Stoneking, M., Kayser, M., Jin, L., and Consortium, T.H.P.-A.S. (2012). Genetic dating indicates that the Asian–Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. Proceedings of the National Academy of Sciences 109, 4574-4579.

24. Lipson, M., Loh, P.R., Patterson, N., Moorjani, P., Ko, Y.C., Stoneking, M., Berger, B., and Reich, D. (2014). Reconstructing Austronesian population history in Island Southeast Asia. Nature Communications 5, 4689.
25. Jinam, T.A., Hong, L.C., Phipps, M.E., Stoneking, M., Ameen, M., Edo, J., and Saitou, N. (2012). Evolutionary history of continental southeast asians: Early train hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. Molecular Biology and Evolution 29, 3513-3527.
26. Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526, 75-81.
27. Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008). Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. Science 319, 1100-1104.
28. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 81, 559-575.
29. Patterson, N., Price, A.L., and Reich, D. (2006). Population Structure and Eigenanalysis. PLoS Genet 2, e190.
30. Alexander, D.H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19, 1655-1664.
31. Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., and François, O. (2014). Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics 196, 973-983.
32. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS genetics 8, e1002967.
33. Chaubey, G., Metspalu, M., Choi, Y., Mägi, R., Romero, I.G., Soares, P., van Oven, M., Behar, D.M., Rootsi, S., Hudjashov, G., et al. (2011). Population Genetic Structure in Indian Austroasiatic Speakers: The Role of Landscape Barriers and Sex-Specific Admixture. Molecular Biology and Evolution 28, 1013-1024.
34. Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. Genetics 193, 1233-1254.
35. Mona, S., Grunz, K.E., Brauer, S., Pakendorf, B., Castr, L., Sudoyo, H., Marzuki, S., Barnes, R.H., Schmidtke, J., Stoneking, M., et al. (2009). Genetic admixture history

of eastern indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. Molecular Biology and Evolution 26, 1865-1877.

36. Anderson, A. (2005). Crossing the Luzon Strait: archaeological chronology in the Batanes Islands, Philippines and the regional sequence of Neolithic dispersal. Journal of Austronesian Studies 1, 25-44.
37. Bulbeck, D. (2008). An integrated perspective on the Austronesian diaspora: The switch from cereal agriculture to maritime foraging in the colonisation of Island Southeast Asia. Australian Archaeology 67, 31-52.
38. Donohue, M., and Denham, T. (2010). Farming and language in Island Southeast Asia: Reframing Austronesian history. Current Anthropology 51, 223-256.
39. Mellars, P., Gori, K.C., Carr, M., Soares, P.A., and Richards, M.B. (2013). Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proceedings of the National Academy of Sciences of the United States of America 110, 10699-10704.
40. Donohue, M., and Denham, T. (in press). Becoming Austronesian: mechanisms of language dispersal across southern Island Southeast Asia. In Austronesian Undressed, D. Gil and J. McWhorter, eds. (Canberra, Pacific Linguistics.
41. Li, H., Wen, B., Chen, S.-J., Su, B., Pramoonjago, P., Liu, Y., Pan, S., Qin, Z., Liu, W., Cheng, X., et al. (2008). Paternal genetic affinity between western Austronesians and Daic populations. BMC Evolutionary Biology 8, 1-12.

CHAPTER 4 – FINAL DISCUSSION AND REMARKS

SEA inherits a unique population history, with distinct interpretations regarding the past demographic events that shaped the region. Excavations in Tabon Cave in Palawan and Callao Cave in Luzon have given the oldest vestiges of the first modern humans so far from the Philippines [200, 274, 275]. These first settlers expanded out of Africa into Australasia following a southern coastal route, most likely facilitated by the prehistoric landmass joining present-day continental and insular SEA [165, 172, 276]. The dispersal of the so-called Australo-Melanesian, or Australoid, Palaeolithic settlers has been confirmed with the presence of basal mtDNA founder haplotypes in the region [243, 253, 258, 277]. Although the arrival of the Upper Palaeolithic settlers is relatively well defined, the following population movements and their impact on the SEA prehistory (in particular in the case of ISEA) are still considerably blurred.

Two major, and partly competing, models have been proposed to explain the population movements in(to) today's ISEA based on genetic (and other) evidences. The most widely accepted model, the "out-of-Taiwan" model, refers to the expansion of Austronesian-speaking populations from Taiwan around 4,500 years. According to this model, the Austronesian agriculturists adopted a Neolithic cultural package including rice, pigs and chickens, which allowed them to colonise and, to some extent replace the hunter-gatherers pre-existing in ISEA [185, 208, 216]. Over recent years, the foundation stones of the "out-of-Taiwan" model have been questioned. Despite the fact that all the Austronesian languages from Madagascar through ISEA and the Pacific, have been assigned to a single 'Malayo-Polynesian' branch, of the same time-depth as the other nine Austronesian branches restricted to Taiwan (placing this region as the homeland) [218, 278], languages can be transmitted horizontally as a part of a cultural shift [214], and so a Taiwanese homeland does not automatically imply a massive agriculturally fuelled mode of dispersal, let alone population replacement. The archaeological evidence also suggests that several elements of the so called "out-of-Taiwan" dispersal were present in ISEA before the arrival of the Austronesian speakers [214]. An alternative model was formulated proposing environmentally related demographic changes in the Late Pleistocene/Early Holocene, before the proposed Neolithic/Austronesian dispersal. At the end of the LGM, sea levels began to rise due to global warming, flooding almost half of the prehistoric Sunda landmass, ultimately resulting in the present-day ISEA topography [165]. These drastic climate changes likely triggered massive autochthonous dispersals. Population movements as a result of postglacial alterations in ISEA have been supported by both archaeological and genetic evidences [172, 253, 264], but do not entirely explain linguistic landscape of the region.

These two very different interpretations of ISEA prehistory appear to be somewhat lacking in complexity. A more comprehensive model, accounting both for the effects of a changing environment and the development new ways of subsistence, needs to be considered. Therefore, this work aimed to perform a comprehensive genetic analysis, combining mtDNA and genome-wide data, to assess the population movements that reflect the genetic landscape of the region. Taking advantage of the secure dating of population movements provided by founder analyses of mtDNA lineages, we performed an extensive analysis of the dispersal times and the contribution of different migrations to present-day ISEA populations. This analysis revealed that the migration in the Early Holocene (~8 ka), contributed to almost half of the lineages of the current ISEA gene pool. Nevertheless, a clear signal of a later migration in the mid-Holocene (~4-5 ka) also detected, both from Taiwan and MSEA, which in total may have contributed to 20-30% of present day ISEA gene pool (Paper I). These two minor mid-Holocene waves most likely represent the Neolithic dispersal from both MSEA and South China, via Taiwan. This initial work confirmed the complex ongoing migration history of ISEA, and more importantly, highlighted (apart from the three major previously studied haplogroups E, B4a1, M7c3c) [172, 248], an additional possible set of rarer mtDNA founder clusters involved in the migratory events of the region. The analysis at high resolution (i.e., the whole-mtDNA sequences) of these remaining rarer mtDNA lineages confirmed the two major past demographic events, the postglacial expansions in the Early Holocene and an "out-of-Taiwan" dispersal in the mid-Holocene, thus cataloguing all maternal lineages of the region (Paper II). The genome-wide analysis, although lacking power in providing a fine time-scale resolution of population movements, also shed light on the complex genetic architecture of the MSEA and ISEA populations. The first appears to share a mixed ancestry between South and East Asian populations, and the latter harbours a very interesting and complex genetic structure, with West ISEA showing ancestral components related to MSEA, and the eastern side of ISEA showing an influx of Papuan-speaking populations (Paper III).

Altogether, the mtDNA and the genome-wide results suggest that the simplistic two-step arrival (at ~50-60 ka and later at ~8 or ~4 ka) model that prevailed for a long time in the field, does not reflect the complex genetic history of the region. In fact, the results place ISEA in the crossroad of multiple prehistoric migrations from continental SEA, Taiwan and the Pacific islands. In summary, this work provides genetic evidence of at least two migratory events at different times, routes and with different consequences for the history region, following the first settlement:

(i) Large-scale migrations into ISEA in the Late Glacial and postglacial periods comprising the Late Pleistocene and Early Holocene:

From this work, the clades that likely expanded in the Early Holocene postglacial period belong to the haplogroups B4a1a and E (Paper I) and haplogroups B5b1c, F3b1, N9a6, B4c1b2a2, R9c1a (Paper II). The postglacial lineages have direct ancestry in mainland Asia at least 15 ka, and lack branching nodes within the time frame of the flooding period (~15–8 ka), which corroborates the catastrophic effects of the climate changes on the effective population size. Moreover, according to their phylogeographic patterns, it seems likely that these lineages evolved within the vicinity of continental Asia, and were caught up in a dramatic series of dispersals and expansions that began in eastern Sundaland/northwest Wallacea, around 10 ka, and ultimately, some of them reached Taiwan in the postglacial period, from both ISEA and South China. The close genetic relationship between MSEA and East Asia is also clearly supported by the genome-wide analysis, where for instance in the PCA, the MSEA populations appear in an intermediate position between (South) China and ISEA in a continuum north-south axis (Paper III).

The most plausible explanation for this scenario is the impact on coastal-dwelling populations of the major environmental changes at the end of the LGM, along with the rising sea level flooding many low-lying areas of the Sunda shelf, which most likely triggered severe population displacements and cultural changes [165, 172, 175, 176, 253]. Archaeological evidence also supports multiple postglacial migrations in the Early Holocene, such as the spread of the Hoabinhian culture across MSEA [279-281] and the spread of a distinct “flake–blade technocomplex”, characterized with flakes detached from rotated multiplatform cores, across ISEA [233, 279, 282].

(ii) Low-scale migrations into ISEA in the mid-to-Late Holocene:

The expansion of rice agriculturalists from South China into ISEA and MSEA, ultimately leading to the spread of the Austronesian and the Austroasiatic languages, respectively, has been considered to be one of the main factors shaping the demographic history of the region. In this work is identified a strong signal of Taiwanese Neolithic

ancestry in ISEA in haplogroups M7c3c and M7b3 from Paper I, and haplogroups B4b1a2, F1a4a, Y2a1 and D5b1c1 from Paper II, that altogether provide the maternal genetic evidence that the expansion from South China into Taiwan and later ISEA, did indeed occur as the archaeological/linguistic model predicts. However, these mid-Holocene lineages only account for ~20% of the current mitochondrial gene pool of ISEA, supporting with a higher confidence (as current information is based on complete mtDNA lineages) previous claims [243]. These findings bring a striking insight into the gene flow history of ISEA, since they imply that the dispersal of the Neolithic culture was not a large-scale demic diffusion as claimed in the "out-of-Taiwan" model [185, 216], but instead it appears to have occurred largely as a process of cultural diffusion and language shift.

As previously mentioned, Taiwan's aboriginal maternal gene pool included not only ancient lineages associated with the Asian continent before sea-level rises separated it from Taiwan, but also a few lineages that arrived from ISEA in the Early Holocene, most probably due to the dramatic postglacial population movements. Taking this in consideration, in theory it is possible that some of ISEA founders might have back-migrated to ISEA in the mid-Holocene. However, no evidence was found of putative mid-Holocene founders within the mitogenome tree of B4a1a, E1, E2 and F3b. Notably, this work confirmed for the first time that all of the lineages that showed a mid-Holocene "out-of-Taiwan" ancestry in ISEA are also Neolithic markers for the settlement of Taiwan from South China, where the putative initial spread of rice-agriculturists across SEA took place. By looking at the genome-wide data of the region, a more ancient single genetic ancestry of ISEA and Taiwan related to continental Asia is also clear. Nevertheless, it also show evidence for gene flow between South China (Dai) to the aboriginal populations of Taiwan, and from those to several Austronesian-speaking populations (Paper III). In this scenario it appears that rice farmers settled in Taiwan from South China around 7–6 ka, and remained somewhat isolated from the autochthonous Taiwanese populations. Later, around ~4-5 ka, these populations spread south carrying with them the Neolithic package (the characteristic red-slipped pottery and domestication practices) and the Proto-Malayo-Polynesian language into ISEA and the Pacific [185, 218, 254].

The results presented here confirms that the population history of ISEA (and SEA as a whole) is much more rich and complex than a simple two-step model. The genetic landscape of ISEA populations is consistent with the occurrence of both migratory events that have been postulated before: the ancestral gene flow from continental Asia (most likely in Late Pleistocene/ Early Holocene) and a more recent small-scale genetic

incursion from Taiwan (most likely representing the spread of the Neolithic practices and Austronesian languages), though each one with different impacts for the populations gene pool. Therefore this work allows us to reconcile the previous, partly competing, interpretations of the history of the region into a new more integrated model. However, although the Y chromosome data presented here (and in the overall literature) revealed a quite similar picture of ISEA, advances in whole Y-chromosome sequencing hold a promising potential for the analysis of the human paternal phylogeny to its maximal resolution (as previously applied to other populations [283]). Therefore, this approach could be important to support or modify this new integrated model, and ultimately provide the definitive picture for the population history of the region. It also it remains to evaluate at a high-resolution level whether a similar clear Chinese signal will be detected in the Y-chromosome pool of lineages proposed as possible markers for mid-Holocene “out-of-Taiwan” migration (haplogroups O1a, O1a2 and O3a). This could provide new perspectives into the cultural processes involved in the spread of agriculture.

Finally, the knowledge of population genetic variation is vital not only for evolutionary studies, but also biomedical studies. In that sense, this work has contributed with substantial data for a more comprehensive view of the population genetic structure of SEA. As for future analyses it would be interesting to use these genome-wide data to evaluate the genomic effects of local demographic and selective pressures across specific regions of SEA. This could potentially provide informative insights into the complex interactions between demographic factors and population-specific selective sweeps. Additionally, further genetic characterization of, for instance, the so-called relict groups of SEA at a genome-wide level, could also be important to illuminate further forces shaping the genetic diversity of human populations.

REFERENCES

1. Jobling, M, Hollox, E, Hurles, M, Kivisild, T, and Tyler-Smith, C, *Human evolutionary genetics, second edition*. 2013, New York: Taylor & Francis Group. 650 pp.

2. Cavalli-Sforza, LL and Feldman, MW, *The application of molecular genetic approaches to the study of human evolution.* Nat Genet, 2003. **33**: 266-75.

3. Keagle, MB, *DNA, chromosomes, and cell division*, in *The Principles of Clinical Cytogenetics*, Gersen, SL and Keagle, MB, Editors. 2013, Springer: New York. p. 9-21.

4. Watson, JD, Baker, TA, Bell, SP, Gann, A, Levine, M, and Losick, R, *Molecular biology of the gene*. 2013, London: Pearson Education. 912 pp.

5. Butler, JM, *Forensic DNA typing: biology, technology, and genetics of STR markers*. 2005, Boston: Elsevier Academic Press. 688 pp.

6. Washietl, S, Pedersen, JS, Korbel, JO, Stocsits, C, Gruber, AR, Hackermüller, J, Hertel, J, Lindemeyer, M, Reiche, K, and Tanzer, A, *Structured RNAs in the ENCODE selected regions of the human genome.* Genome Res, 2007. **17**(6): 852-64.

7. Weinstock, GM, *ENCODE: more genomic empowerment.* Genome Res, 2007. **17**(6): 667-68.

8. Mercer, TR, Gerhardt, DJ, Dinger, ME, Crawford, J, Trapnell, C, Jeddeloh, JA, Mattick, JS, and Rinn, JL, *Targeted RNA sequencing reveals the deep complexity of the human transcriptome.* Nat Biotechnol, 2012. **30**(1): 99-104.

9. Mattick, JS, *The genetic signatures of noncoding RNAs.* PLoS Genet, 2009. **5**(4): e1000459.

10. Hardison, RC, *Conserved noncoding sequences are reliable guides to regulatory elements.* Trends Genet, 2000. **16**(9): 369-72.

11. Bergman, CM and Kreitman, M, *Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences.* Genome Res, 2001. **11**(8): 1335-45.

12. Shabalina, SA, Ogurtsov, AY, Kondrashov, VA, and Kondrashov, AS, *Selective constraint in intergenic regions of human and mouse genomes.* Trends Genet, 2001. **17**(7): 373-76.

13. Strachan, T, Goodship, J, and Chinnery, P, *Genetics and genomics in Medicine*. 2014, New York: Taylor & Francis Group. 500 pp.

14. Jobling, MA and Gill, P, *Encoded evidence: DNA in forensic analysis.* Nat Rev Genet, 2004. **5**(10): 739-51.

15. Graur, D, *Single-base mutation.* eLS, 2008.

16. Griffiths, AJF, *Modern genetic analysis: integrating genes and genomes*. 2002, New York: W. H. Freeman. 736 pp.

17. Akey, JM, Eberle, MA, Rieder, MJ, Carlson, CS, Shriver, MD, Nickerson, DA, and Kruglyak, L, *Population history and natural selection shape patterns of genetic variation in 132 genes.* PLoS Biol, 2004. **2**(10): e286.

18. Nachman, MW and Crowell, SL, *Estimate of the mutation rate per nucleotide in humans.* Genetics, 2000. **156**(1): 297-304.

19. Kong, A, Frigge, ML, Masson, G, Besenbacher, S, Sulem, P, Magnusson, G, Gudjonsson, SA, Sigurdsson, A, Jonasdottir, A, and Jonasdottir, A, *Rate of de*

novo mutations and the importance of father's age to disease risk. Nature, 2012. **488**(7412): 471-75.

20. Weber, JL and Wong, C, *Mutation of human short tandem repeats.* Hum Mol Genet, 1993. **2**(8): 1123-28.

21. Sun, JX, Helgason, A, Masson, G, Ebenesersdottir, SS, Li, H, Mallick, S, Gnerre, S, Patterson, N, Kong, A, Reich, D, and Stefansson, K, *A direct characterization of human mutation based on microsatellites.* Nat Genet, 2012. **44**(10): 1161-65.

22. Pierce, BA, *Genetics: a conceptual approach*. 2012, New York: W. H. Freeman. 400 pp.

23. Pengelly, RJ, Tapper, W, Gibson, J, Knut, M, Tearle, R, Collins, A, and Ennis, S, *Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations.* BMC Genomics, 2015. **16**(1): 666.

24. Novembre, J and Ramachandran, S, *Perspectives on human population structure at the cusp of the sequencing era.* Annu Rev Genom Hum G, 2011. **12**(1): 245-74.

25. Hedrick, PW, *Genetics of populations*. 2011, Sudbury: Jones & Bartlett Learning, LLC. 675 pp.

26. Lemey, P, Salemi, M, and Vandamme, AM, *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. 2009, Cambridge: Cambridge University Press. 723 pp.

27. Vitti, JJ, Grossman, SR, and Sabeti, PC, *Detecting natural selection in genomic data.* Annu Rev Genet, 2013. **47**(1): 97-120.

28. Nielsen, R, Hellmann, I, Hubisz, M, Bustamante, C, and Clark, AG, *Recent and ongoing selection in the human genome.* Nat Rev Genet, 2007. **8**(11): 857-68.

29. Vasseur, E and Quintana-Murci, L, *The impact of natural selection on health and disease: uses of the population genetics approach in humans.* Evol Appl, 2013. **6**(4): 596-607.

30. Sabeti, P, Schaffner, S, Fry, B, Lohmueller, J, Varilly, P, Shamovsky, O, Palma, A, Mikkelsen, T, Altshuler, D, and Lander, E, *Positive natural selection in the human lineage.* Science, 2006. **312**(5780): 1614-20.

31. Hurst, LD, *Genetics and the understanding of selection.* Nat Rev Genet, 2009. **10**(2): 83-93.

32. Kimura, M, *Evolutionary rate at the molecular level.* Nature, 1968. **217**(5129): 624-26.

33. Kimura, M, *The neutral theory of molecular evolution*. 1984, Cambridge: Cambridge University Press. 367 pp.

34. Zuckerkandl, E and Pauling, L, *Evolutionary divergence and convergence in proteins*, in *Evolving genes and proteins*, Bryson V and HJ, V, Editors. 1965, Academic Press. p. 97-166.

35. Yang, Z and Rannala, B, *Molecular phylogenetics: principles and practice.* Nat Rev Genet, 2012. **13**(5): 303-14.

36. Kumar, S, *Molecular clocks: four decades of evolution.* Nat Rev Genet, 2005. **6**(8): 654-62.

37. Holsinger, KE and Weir, BS, *Genetics in geographically structured populations: defining, estimating and interpreting FST.* Nat Rev Genet, 2009. **10**(9): 639-50.

38. Reddy, BN, *Basics for the construction of phylogenetic trees.* Webmedcentral, 2011. **2**(12): WMC00256.

39. Jill Harrison, C and Langdale, JA, *A step by step guide to phylogeny reconstruction.* Plant J, 2006. **45**(4): 561-72.

40. Sokal, RR, *A statistical method for evaluating systematic relationships.* Univ Kans Sci Bull, 1958. **38**: 1409-38.

41. Saitou, N and Nei, M, *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol, 1987. **4**(4): 406-25.

42. Rzhetsky, A and Nei, M, *A simple method for estimating and testing minimum-evolution trees.* Mol Biol Evol, 1992. **9**(5): 945-67.

43. Sleator, RD, *A beginner's guide to phylogenetics.* Microb Ecol, 2013. **66**(1): 1-4.

44. Richards, MB, Macaulay, VA, Bandelt, HJ, and Sykes, BC, *Phylogeography of mitochondrial DNA in western Europe.* Ann Hum Genet, 1998. **62**(3): 241-60.

45. Bandelt, H-J, Quintana-Murci, L, Salas, A, and Macaulay, V, *The fingerprint of phantom mutations in mitochondrial DNA data.* Am J Hum Genet, 2002. **71**(5): 1150-60.

46. Bandelt, H, Forster, P, and Rohl, A, *Median-joining networks for inferring intraspecific phylogenies.* Mol Biol Evol, 1999. **16**: 37 - 48.

47. Bandelt, H-J, Kong, Q-P, Richards, M, and Macaulay, V, *Estimation of mutation rates and coalescence times: some caveats*, in *Human mitochondrial DNA and the evolution of Homo sapiens*, Bandelt, H-J, Macaulay, V, and Richards, M, Editors. 2006, Springer: New York. p. 47-90.

48. Jukes, TH and Cantor, CR, *Evolution of protein molecules*, in *Mammalian protein metabolism*, Munro, HN, Editor. 1969, Academic Press: New York. p. 21-132.

49. Kimura, M, *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* J Mol Evol, 1980. **16**(2): 111-20.

50. Hasegawa, M, Kishino, H, and Yano, T-a, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.* J Mol Evol, 1985. **22**(2): 160-74.

51. Tamura, K and Nei, M, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.* Mol Biol Evol, 1993. **10**(3): 512-26.

52. Tavaré, S, *Some probabilistic and statistical problems in the analysis of DNA sequences.* Lectures on mathematics in the life sciences, 1986. **17**: 57-86.

53. Waddell, PJ and Steel, MA, *General time-reversible distances with unequal rates across sites: mixing Γ and inverse Gaussian distributions with invariant sites.* Mol Phylogenet Evol, 1997. **8**(3): 398-414.

54. Kingman, JF, *Origins of the coalescent: 1974-1982.* Genetics, 2000. **156**(4): 1461-63.

55. Forster, P, Harding, R, Torroni, A, and Bandelt, H, *Origin and evolution of Native American mtDNA variation: a reappraisal.* Am J Hum Genet, 1996. **59**: 935 - 45.

56. Stumpf, MP and Goldstein, DB, *Genealogical and evolutionary inference with the human Y chromosome.* Science, 2001. **291**(5509): 1738-42.

57. Rutschmann, F, *Molecular dating of phylogenetic trees: a brief review of current methods that estimate divergence times.* Divers Distrib, 2006. **12**(1): 35-48.

58. Drummond, AJ, Ho, SY, Phillips, MJ, and Rambaut, A, *Relaxed phylogenetics and dating with confidence.* PLoS Biol, 2006. **4**(5): e88.

59. Richards, M, Macaulay, V, Hickey, E, Vega, E, Sykes, B, Guida, V, Rengo, C, Sellitto, D, Cruciani, F, Kivisild, T, Villems, R, Thomas, M, Rychkov, S, Rychkov, O, Rychkov, Y, Golge, M, Dimitrov, D, Hill, E, Bradley, D, Romano, V, Cali, F, Vona, G, Demaine, A, Papiha, S, Triantaphyllidis, C, Stefanescu, G, Hatina, J, Belledi, M, Di Rienzo, A, Novelletto, A, Oppenheim, A, Norby, S, Al-Zaheri, N, Santachiara-Benerecetti, S, Scozari, R, Torroni, A, and Bandelt, HJ, *Tracing European founder lineages in the Near Eastern mtDNA pool.* Am J Hum Genet, 2000. **67**(5): 1251-76.

60. Richards, MB and Macaulay, V, *Genetic data and the colonization of Europe: genealogies and founders*, in *Archaeogenetics: DNA and the population prehistory of Europe*, Renfrew, C and Boyle, K, Editors. 2000, McDonald Institute for Archaelogical Research: Cambridge. p. 139-51.

61. Barbujani, G, Bertorelle, G, and Chikhi, L, *Evidence for Paleolithic and Neolithic gene flow in Europe.* Am J Hum Genet, 1998. **62**(2): 488-92.

62. Richards, M and Sykes, B, *Reply to Barbujani et al.* Am J Hum Genet, 1998. **62**(2): 491-92.

63. Karydas, CG, Gitas, IZ, Koutsogiannaki, E, Lydakis-Simantiris, N, and Silleos, G, *Evaluation of spatial interpolation techniques for mapping agricultural topsoil properties in Crete.* EARSeL eProceedings, 2009. **8**(1): 26-39.

64. Yang, C-S, Kao, S-P, Lee, F-B, and Hung, P-S, *Twelve different interpolation methods: A case study of Surfer 8.0*. Vol. 35. 2004, Istanbul, Turkey: Abstr Vol Geo-Imagery Bridging Continents, XXth ISPRS Congress. 778-85 pp.

65. Loh, P-R, Lipson, M, Patterson, N, Moorjani, P, Pickrell, JK, Reich, D, and Berger, B, *Inferring admixture histories of human populations using linkage disequilibrium.* Genetics, 2013. **193**(4): 1233-54.

66. Engelhardt, BE and Stephens, M, *Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis.* PLoS Genet, 2010. **6**(9): e1001117.

67. Pritchard, JK, Stephens, M, and Donnelly, P, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): 945-59.

68. Patterson, N, Price, AL, and Reich, D, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): e190.

69. Alexander, DH, Novembre, J, and Lange, K, *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res, 2009. **19**(9): 1655-64.

70. Tang, H, Peng, J, Wang, P, and Risch, NJ, *Estimation of individual admixture: analytical and study design considerations.* Genet Epidemiol, 2005. **28**(4): 289-301.

71. Liu, Y, Nyunoya, T, Leng, S, Belinsky, SA, Tesfaigzi, Y, and Bruse, S, *Softwares and methods for estimating genetic ancestry in human populations.* Hum Genomics, 2013. **7**(1): 1.

72. Reich, D, Price, AL, and Patterson, N, *Principal component analysis of genetic data.* Nat Genet, 2008. **40**(5): 491-92.

73. Sankararaman, S, Kimmel, G, Halperin, E, and Jordan, MI, *On the inference of ancestries in admixed populations.* Genome Res, 2008. **18**(4): 668-75.

74. Maples, BK, Gravel, S, Kenny, EE, and Bustamante, CD, *RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference.* Am J Hum Genet, 2013. **93**(2): 278-88.

75. Patterson, N, Moorjani, P, Luo, Y, Mallick, S, Rohland, N, Zhan, Y, Genschoreck, T, Webster, T, and Reich, D, *Ancient admixture in human history.* Genetics, 2012. **192**(3): 1065-93.

76. Jorde, LB, Watkins, WS, and Bamshad, MJ, *Population genomics: a bridge from evolutionary history to genetic medicine.* Hum Mol Genet, 2001. **10**(20): 2199-207.

77. Tishkoff, SA and Verrelli, BC, *Patterns of human genetic diversity: implications for human evolutionary history and disease.* Annu Rev Genom Hum G, 2003. **4**(1): 293-340.

78. Underhill, PA and Kivisild, T, *Use of y chromosome and mitochondrial DNA population structure in tracing human migrations.* Annu Rev Genet, 2007. **41**: 539-64.

79. Pakendorf, B and Stoneking, M, *Mitochondrial DNA and human evolution.* Annu Rev Genom Hum G, 2005. **6**(1): 165-83.

80. Strachan, T and Read, A, *Human Molecular Genetics*. 2010: Taylor & Francis Group. 812 pp.

81. Gray, MW, Burger, G, and Lang, BF, *The origin and early evolution of mitochondria.* Genome Biol, 2001. **2**(6): 1018.1-18.5.

82. Anderson, S, Bankier, AT, Barrell, BG, De Bruijn, M, Coulson, AR, Drouin, J, Eperon, I, Nierlich, D, Roe, BA, and Sanger, F, *Sequence and organization of the human mitochondrial genome.* Nature, 1981. **290**(5806): 457-65.

83. Andrews, RM, Kubacka, I, Chinnery, PF, Lightowlers, RN, Turnbull, DM, and Howell, N, *Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA.* Nat Genet, 1999. **23**(2): 147-47.

84. Heyer, E, Zietkiewicz, E, Rochowski, A, Yotova, V, Puymirat, J, and Labuda, D, *Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees.* Am J Hum Genet, 2001. **69**(5): 1113-26.

85. Thomas, J, David, S, and Kevin, S, *A high observed substitution rate in the human mitochondrial DNA control region.* Nat Genet, 1997. **15**: 363-68.

86. Mishmar, D, Ruiz-Pesini, E, Golik, P, Macaulay, V, Clark, AG, Hosseini, S, Brandon, M, Easley, K, Chen, E, and Brown, MD, *Natural selection shaped regional mtDNA variation in humans.* PNAS, 2003. **100**(1): 171-76.

87. Kivisild, T, Shen, P, Wall, DP, Do, B, Sung, R, Davis, K, Passarino, G, Underhill, PA, Scharfe, C, Torroni, A, Scozzari, R, Modiano, D, Coppa, A, de Knijff, P, Feldman, M, Cavalli-Sforza, LL, and Oefner, PJ, *The role of selection in the evolution of human mitochondrial genomes.* Genetics, 2006. **172**(1): 373-87.

88. Endicott, P, Ho, SYW, Metspalu, M, and Stringer, C, *Evaluating the mitochondrial timescale of human evolution.* Trends Ecol Evol, 2009. **24**(9): 515-21.

89. Pereira, L, Silva, NM, Franco-Duarte, R, Fernandes, V, Pereira, JB, Costa, MD, Martins, H, Soares, P, Behar, DM, Richards, MB, and Macaulay, V, *Population*

expansion in the North African late Pleistocene signalled by mitochondrial DNA haplogroup U6. BMC Evol Biol, 2010. **10**: 390.

90. Soares, P, Alshamali, F, Pereira, JB, Fernandes, V, Silva, NM, Afonso, C, Costa, MD, Musilova, E, Macaulay, V, Richards, MB, Cerny, V, and Pereira, L, *The expansion of mtDNA haplogroup L3 within and out of Africa.* Mol Biol Evol, 2012. **29**(3): 915-27.

91. Soares, P, Ermini, L, Thomson, N, Mormina, M, Rito, T, Röhl, A, Salas, A, Oppenheimer, S, Macaulay, V, and Richards, MB, *Correcting for purifying selection: an improved human mitochondrial molecular clock.* Am J Hum Genet, 2009. **84**(6): 740-59.

92. Sutovsky, P, Moreno, RD, Ramalho-Santos, J, Dominko, T, Simerly, C, and Schatten, G, *Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos.* Biol Reprod, 2000. **63**(2): 582-90.

93. Schwartz, M and Vissing, J, *Paternal inheritance of mitochondrial DNA.* New Engl J Med, 2002. **347**(8): 576-80.

94. Torroni, A, Sukernik, RI, Schurr, TG, Starikorskaya, Y, Cabell, MF, Crawford, MH, Comuzzie, AG, and Wallace, DC, *mtDNA variation of aboriginal Siberians reveals distinct genetic affinities with Native Americans.* Am J Hum Genet, 1993. **53**(3): 591-608.

95. Cann, RL, *DNA and human origins.* Annu Rev Anthropol, 1988. **17**: 127-43.

96. Cerezo, M, Achilli, A, Olivieri, A, Perego, UA, Gómez-Carballa, A, Brisighelli, F, Lancioni, H, Woodward, SR, López-Soto, M, and Carracedo, Á, *Reconstructing ancient mitochondrial DNA links between Africa and Europe.* Genome Res, 2012. **22**(5): 821-26.

97. Kivisild, T, *Maternal ancestry and population history from whole mitochondrial genomes.* Investig Genet, 2015. **6**(1): 3.

98. van Oven, M and Kayser, M, *Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation.* Hum Mutat, 2009. **30**(2): E386 - E94.

99. Maca-Meyer, N, Gonzalez, AM, Larruga, JM, Flores, C, and Cabrera, VM, *Major genomic mitochondrial lineages delineate early human expansions.* BMC Genet, 2001. **2**: 13.

100. Oppenheimer, S, *Out of Eden*. 2004, London: Constable and Robison Ltd. 440 pp.

101. Richards, M, Bandelt, HJ, Kivisild, T, and Oppenheimer, S, *A model for the dispersal of modern humans out of Africa*, in *Mitochondrial DNA and the evolution of Homo sapiens.*, Bandelt, HJ, Macaulay, V, and Richards, M, Editors. 2006, Springer–Verlag: Berlin. p. 225–65.

102. Shea, JJ, *Transitions or turnovers? Climatically-forced extinctions of Homo sapiens and Neanderthals in the east Mediterranean Levant.* Quat Sci Rev, 2008. **27**(23–24): 2253-70.

103. Torroni, A, Achilli, A, Macaulay, V, Richards, M, and Bandelt, H, *Harvesting the fruit of the human mtDNA tree.* Trends Genet, 2006. **22**(6): 339-45.

104. Friedlaender, J, Schurr, T, Gentz, F, Koki, G, Friedlaender, F, Horvat, G, Babb, P, Cerchio, S, Kaestle, F, Schanfield, M, Deka, R, Yanagihara, R, and Merriwether, DA, *Expanding Southwest Pacific mitochondrial haplogroups P and Q.* Mol Biol Evol, 2005. **22**: 1506-17.

105. Ingman, M and Gyllensten, U, *Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines.* Genome Res, 2003. **13**: 1600-06.

106. Kong, Q-P, Bandelt, H-J, Sun, C, Yao, Y-G, Salas, A, Achilli, A, Wang, C-Y, Zhong, L, Zhu, C-L, Wu, S-F, Torroni, A, and Zhang, Y-P, *Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations.* Hum Mol Genet, 2006. **15**: 2076-86.

107. Kong, Q-P, Yao, Y-G, Sun, C, Bandelt, H-J, Zhu, C-L, and Zhang, Y-P, *Phylogeny of East Asian mitochondrial DNA linerages inferred from complete sequences.* Am J Hum Genet, 2003. **73**: 671-76.

108. Merriwether, DA, Hodgson, JA, Friedlaender, FR, Allaby, R, Cerchio, S, Koki, G, and Friedlaender, JS, *Ancient mitochondrial M haplogroups identified in the Southwest Pacific.* PNAS, 2005. **102**: 13034-39.

109. Palanichamy, M, Sun, C, Agrawa, S, Bandelt, H-J, Kong, Q-P, Khan, F, Wang, C-Y, Chaudhuri, TP, Palla, V, and Zhang, Y-P, *Phylogeny of mtDNA macrohaplogroup N in India based on complete sequencing: implications for the peopling of South Asia.* Am J Hum Genet, 2004. **75**: 966-78.

110. Pellekaan, SMVH, Ingman, M, Roberts-Thomson, J, and Harding, RM, *Mitochondrial genomics identifies major haplogroups in Aboriginal Australians.* AmJ Physic Anthropol, 2006. **131**: 282-94.

111. Sun, C, Kong, Q-P, Palanichamy, M, Agrawal, S, Bandelt, H-J, Yao, Y-G, Khan, F, Zhu, C-L, Chaudhuri, TK, and Zhang, Y-P, *The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes.* Mol Biol and Evol, 2006. **23**: 683–90.

112. Thangaraj, K, Chaubey, G, Kivisild, T, Reddy, AG, Singh, VK, Rasalkar, AA, and Singh, L, *Reconstructing the origin of Andaman Islanders.* Science, 2005. **308**: 996.

113. Thangaraj, K, Chaubey, G, Singh, VK, Thanseem, I, Reddy, AG, and Singh, L, *In situ origin of deep rooting lineages of mitochondrial Macrohaplogroup 'M' in India.* BMC Genomics, 2006. **15**: 151.

114. Macaulay, V, Hill, C, Achilli, A, Rengo, C, Clarke, D, Meehan, W, Blackburn, J, Semino, O, Scozzari, R, Cruciani, F, Taha, A, Shaari, NK, Raja, JM, Ismail, P, Zainuddin, Z, Goodwin, W, Bulbeck, D, Bandelt, H-J, Oppenheimer, S, Torroni, A, and Richards, M, *Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Human Mitochondrial Genomes.* Science, 2005. **308**: 1034-36.

115. Pereira, L, Richards, M, Goios, A, Alonso, A, Albarrán, C, Garcia, O, Behar, DM, Gölge, M, Hatina, Ji, Al-Gazali, L, Bradley, DG, Macaulay, V, and Amorim, A, *High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium.* Genome Res, 2005. **15**: 19–24.

116. Torroni, A, Bandelt, HJ, Macaulay, V, Richards, M, Cruciani, F, Rengo, C, Martinez-Cabrera, V, Villems, R, Kivisild, T, Metspalu, E, Parik, J, Tolk, HV, Tambets, K, Forster, P, Karger, B, Francalacci, P, Rudan, P, Janicijevic, B, Rickards, O, Savontaus, ML, Huoponen, K, Laitinen, V, Koivumaki, S, Sykes, B, Hickey, E, Novelletto, A, Moral, P, Sellitto, D, Coppa, A, Al-Zaheri, N, Santachiara-Benerecetti, AS, Semino, O, and Scozzari, R, *A signal, from human mtDNA, of postglacial recolonization in Europe.* Am J Hum Genet, 2001. **69**(4): 844-52.

117. Pala, M, Olivieri, A, Achilli, A, Accetturo, M, Metspalu, E, Reidla, M, Tamm, E, Karmin, M, Reisberg, T, Hooshiar Kashani, B, Perego, UA, Carossa, V, Gandini, F, Pereira, JB, Soares, P, Angerhofer, N, Rychkov, S, Al-Zahery, N, Carelli, V, Sanati, MH, Houshmand, M, Hatina, J, Macaulay, V, Pereira, L, Woodward, SR, Davies, W, Gamble, C, Baird, D, Semino, O, Villems, R, Torroni, A, and Richards, MB, *Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia.* Am J Hum Genet, 2012. **90**(5): 915-24.

118. Soares, P, Achilli, A, Semino, O, Davies, W, Macaulay, V, Bandelt, H-J, Torroni, A, and Richards, MB, *The archaeogenetics of Europe.* Curr Biol, 2010. **20**(4): R174-R83.

119. Pala, M, Achilli, A, Olivieri, A, Kashani, BH, Perego, UA, Sanna, D, Metspalu, E, Tambets, K, Tamm, E, Accetturo, M, Carossa, V, Lancioni, H, Panara, F, Zimmermann, B, Huber, G, Al-Zahery, N, Brisighelli, F, Woodward, SR, Francalacci, P, Parson, W, Salas, A, Behar, DM, Villems, R, Semino, O, Bandelt, H-J, and Torroni, A, *Mitochondrial haplogroup U5b3: a distant echo of the epipaleolithic in Italy and the legacy of the early Sardinians.* Am J Hum Genet, 2009. **84**(6): 814-21.

120. Malyarchuk, B, Derenko, M, Grzybowski, T, Perkova, M, Rogalla, U, Vanecek, T, and Tsybovsky, I, *The peopling of Europe from the mitochondrial haplogroup U5 perspective.* PLoS One, 2010. **5**(4): e10285.

121. Olivieri, A, Pala, M, Gandini, F, Kashani, BH, Perego, UA, Woodward, SR, Grugni, V, Battaglia, V, Semino, O, Achilli, A, Richards, MB, and Torroni, A, *Mitogenomes from Two Uncommon Haplogroups Mark Late Glacial/Postglacial Expansions from the Near East and Neolithic Dispersals within Europe.* PLoS One, 2013. **8**(7): e70492.

122. Brotherton, P, Haak, W, Templeton, J, Brandt, G, Soubrier, J, Jane Adler, C, Richards, SM, Der Sarkissian, C, Ganslmeier, R, Friederich, S, Dresely, V, van Oven, M, Kenyon, R, Van der Hoek, MB, Korlach, J, Luong, K, Ho, SYW, Quintana-Murci, L, Behar, DM, Meller, H, Alt, KW, Cooper, A, and The Genographic, C, *Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans.* Nat Commun, 2013. **4**: 1764-64.

123. Torroni, A, Schurr, TG, Cabell, MF, Brown, MD, Neel, JV, Larsen, M, Smith, DG, Vullo, CM, and Wallace, DC, *Asian affinities and continental radiation of the four founding Native American mtDNAs.* Am J Hum Genet, 1993. **53**(3): 563-90.

124. Skaletsky, H, Kuroda-Kawaguchi, T, Minx, PJ, Cordum, HS, Hillier, L, Brown, LG, Repping, S, Pyntikova, T, Ali, J, and Bieri, T, *The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes.* Nature, 2003. **423**(6942): 825-37.

125. Mangs, AH and Morris, BJ, *The human pseudoautosomal region (PAR): origin, function and future.* Curr Genomics, 2007. **8**(2): 129-36.

126. Hughes, JF and Rozen, S, *Genomics and genetics of human and primate Y chromosomes.* Annu Rev Genom Hum G, 2012. **13**(1): 83-108.

127. Kayser, M, Roewer, L, Hedman, M, Henke, L, Henke, J, Brauer, S, Krüger, C, Krawczak, M, Nagy, M, and Dobosz, T, *Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs.* Am J Hum Genet, 2000. **66**(5): 1580-88.

128. Karafet, T, Mendez, F, Meilerman, M, Underhill, P, Zegura, S, and Hammer, M, *New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree.* Genome Res, 2008. **18**(5): 830 - 38.

129. Onofri, V, Alessandrini, F, Turchi, C, Pesaresi, M, Buscemi, L, and Tagliabracci, A, *Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs.* Forensic Sci Int, 2006. **157**(1): 23-35.

130. Forster, P, Röhl, A, Lünnemann, P, Brinkmann, C, Zerjal, T, Tyler-Smith, C, and Brinkmann, B, *A short tandem repeat–based phylogeny for the human Y chromosome.* Am J Hum Genet, 2000. **67**(1): 182-96.

131. Zhivotovsky, LA, Underhill, PA, and Feldman, MW, *Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size.* Mol Biol Evol, 2006. **23**(12): 2268-70.

132. Consortium, YC, *A nomenclature system for the tree of human Y-chromosomal binary haplogroups.* Genome Res, 2002. **12**(2): 339-48.

133. Underhill, P, Shen, P, Lin, A, Jin, L, Passarino, G, Yang, W, Kauffman, E, Bonne-Tamir, B, Bertranpetit, J, and Francalacci, P, *Y chromosome sequence variation and the history of human populations.* Nat Genet, 2000. **26**(3): 358 - 61.

134. Jobling, MA and Tyler-Smith, C, *The human Y chromosome: an evolutionary marker comes of age.* Nat Rev Genet, 2003. **4**(8): 598-612.

135. Sengupta, S, Zhivotovsky, LA, King, R, Mehdi, S, Edmonds, CA, Chow, C-ET, Lin, AA, Mitra, M, Sil, SK, and Ramesh, A, *Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists.* Am J Hum Genet, 2006. **78**(2): 202-21.

136. Underhill, PA, Myres, NM, Rootsi, S, Metspalu, M, Zhivotovsky, LA, King, RJ, Lin, AA, Chow, CE, Semino, O, Battaglia, V, Kutuev, I, Jarve, M, Chaubey, G, Ayub, Q, Mohyuddin, A, Mehdi, SQ, Sengupta, S, Rogaev, EI, Khusnutdinova, EK, Pshenichnov, A, Balanovsky, O, Balanovska, E, Jeran, N, Augustin, DH, Baldovic, M, Herrera, RJ, Thangaraj, K, Singh, V, Singh, L, Majumder, P, Rudan, P, Primorac, D, Villems, R, and Kivisild, T, *Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a.* Eur J Hum Genet, 2010. **18**(4): 479-84.

137. Rootsi, S, Zhivotovsky, LA, Baldovic, M, Kayser, M, Kutuev, IA, Khusainova, R, Bermisheva, MA, Gubina, M, Fedorova, SA, Ilumae, AM, Khusnutdinova, EK, Voevoda, MI, Osipova, LP, Stoneking, M, Lin, AA, Ferak, V, Parik, J, Kivisild, T, Underhill, PA, and Villems, R, *A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe.* Eur J Hum Genet, 2007. **15**(2): 204-11.

138. Ebersberger, I, Metzler, D, Schwarz, C, and Pääbo, S, *Genomewide comparison of DNA sequences between humans and chimpanzees.* Am J Hum Genet, 2002. **70**(6): 1490-97.

139. Consortium, IHGS, *Finishing the euchromatic sequence of the human genome.* Nature, 2004. **431**(7011): 931-45.

140. Watson, JD, *The human genome project: past, present, and future.* Science, 1990. **248**(4951): 44-49.

141. Lander, ES, Linton, LM, Birren, B, Nusbaum, C, Zody, MC, Baldwin, J, Devon, K, Dewar, K, Doyle, M, and FitzHugh, W, *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): 860-921.

142. Foster, MW, *Human Genome Diversity Project (HGDP).* eLS, 2003.

143. Consortium, TIH, *A haplotype map of the human genome.* Nature, 2005. **437**(7063): 1299-320.

144. Cavalli-Sforza, LL, *The human genome diversity project: past, present and future.* Nat Rev Genet, 2005. **6**(4): 333-40.

145. Wall, JD and Pritchard, JK, *Assessing the performance of the haplotype block model of linkage disequilibrium.* Am J Hum Genet, 2003. **73**(3): 502-15.

146. Consortium, IH, *Integrating common and rare genetic variation in diverse human populations.* Nature, 2010. **467**(7311): 52-58.

147. The ENCODE Project Consortium, *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.* Nature, 2007. **447**(7146): 799-816.

148. The 1000 Genomes Project Consortium, *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): 56-65.

149. The 1000 Genomes Project Consortium, *A global reference for human genetic variation.* Nature, 2015. **526**(7571): 68-74.

150. Wilkinson, S, Wiener, P, Archibald, AL, Law, A, Schnabel, RD, McKay, SD, Taylor, JF, and Ogden, R, *Evaluation of approaches for identifying population informative markers from high density SNP Chips.* BMC Genet, 2011. **12**(1): 45.

151. Manolio, TA, Brooks, LD, and Collins, FS, *A HapMap harvest of insights into the genetics of common disease.* J Clin Invest, 2008. **118**(5): 1590.

152. Bowcock, A, Ruiz-Linares, A, Tomfohrde, J, Minch, E, Kidd, J, and Cavalli-Sforza, LL, *High resolution of human evolutionary trees with polymorphic microsatellites.* Nature, 1994. **368**(6470): 455-57.

153. Tishkoff, SA, Reed, FA, Friedlaender, FR, Ehret, C, Ranciaro, A, Froment, A, Hirbo, JB, Awomoyi, AA, Bodo, J-M, and Doumbo, O, *The genetic structure and history of Africans and African Americans.* Science, 2009. **324**(5930): 1035-44.

154. Friedlaender, JS, Friedlaender, FR, Reed, FA, Kidd, KK, Kidd, JR, Chambers, GK, Lea, RA, Loo, JH, Koki, G, Hodgson, JA, Merriwether, DA, and Weber, JL, *The genetic structure of Pacific Islanders.* PLoS Genet, 2008. **4**(1): e19.

155. Kopelman, NM, Stone, L, Wang, C, Gefel, D, Feldman, MW, Hillel, J, and Rosenberg, NA, *Genomic microsatellites identify shared Jewish ancestry intermediate between Middle Eastern and European populations.* BMC Genet, 2009. **10**(1): 80.

156. Pugach, I and Stoneking, M, *Genome-wide insights into the genetic history of human populations.* Investig Genet, 2015. **6**(1): 6.

157. Abdulla, MA, Ahmed, I, Assawamakin, A, Bhak, J, Brahmachari, SK, Calacal, GC, Chaurasia, A, Chen, CH, Chen, J, Chen, YT, Chu, J, Cutiongco-de la Paz, EMC, De Ungria, MCA, Delfin, FC, Edo, J, Fuchareon, S, Ghang, H, Gojobori, T, Han, J, Ho, SF, Hoh, BP, Huang, W, Inoko, H, Jha, P, Jinam, TA, Jin, L, Jung, J, Kangwanpong, D, Kampuansai, J, Kennedy, GC, Khurana, P, Kim, HL, Kim, K, Kim, S, Kim, WY, Kimm, K, Kimura, R, Koike, T, Kulawonganunchai, S, Kumar, V, Lai, PS, Lee, JY, Lee, S, Liu, ET, Majumder, PP, Mandapati, KK, Marzuki, S,

Mitchell, W, Mukerji, M, Naritomi, K, Ngamphiw, C, Niikawa, N, Nishida, N, Oh, B, Oh, S, Ohashi, J, Oka, A, Ong, R, Padilla, CD, Palittapongarnpim, P, Perdigon, HB, Phipps, ME, Png, E, Sakaki, Y, Salvador, JM, Sandraling, Y, Scaria, V, Seielstad, M, Sidek, MR, Sinha, A, Srikummool, M, Sudoyo, H, Sugano, S, Suryadi, H, Suzuki, Y, Tabbada, KA, Tan, A, Tokunaga, K, Tongsima, S, Villamor, LP, Wang, E, Wang, Y, Wang, H, Wu, JY, Xiao, H, Xu, S, Yang, JO, Shugart, YY, Yoo, HS, Yuan, W, Zhao, G, and Zilfalil, BA, *Mapping human genetic diversity in Asia.* Science, 2009. **326**(5959): 1541-45.

158. Jin, W, Xu, S, Wang, H, Yu, Y, Shen, Y, Wu, B, and Jin, L, *Genome-wide detection of natural selection in African Americans pre- and post-admixture.* Genome Res, 2012. **22**(3): 519-27.

159. Lupski, JR, Belmont, JW, Boerwinkle, E, and Gibbs, RA, *Clan genomics and the complex architecture of human disease.* Cell, 2011. **147**(1): 32-43.

160. Gonzaga-Jauregui, C, Lupski, JR, and Gibbs, RA, *Human genome sequencing in health and disease.* Annu Rev Med, 2012. **63**: 35-61.

161. Li, JZ, Absher, DM, Tang, H, Southwick, AM, Casto, AM, Ramachandran, S, Cann, HM, Barsh, GS, Feldman, M, Cavalli-Sforza, LL, and Myers, RM, *Worldwide human relationships inferred from genome-wide patterns of variation.* Science, 2008. **319**(5866): 1100-4.

162. Novembre, J, Johnson, T, Bryc, K, Kutalik, Z, Boyko, AR, Auton, A, Indap, A, King, KS, Bergmann, S, and Nelson, MR, *Genes mirror geography within Europe.* Nature, 2008. **456**(7218): 98-101.

163. Bellwood, P, *Prehistory of the Indo-Malaysian Archipelago*. 1997: University of Hawai'i Press Bookspp.

164. Wallace, AR, *On the zoological geography of the Malay Archipelago.* Zoll J Linn Soc-Lond, 1860. **4**(16): 172-84.

165. Oppenheimer, S, *Eden in the east: the drowned continent of Southeast Asia*. 1998, London: Phoenix. 560 pp.

166. Palmer, E, *Out of Sunda? Provenance of the Jōmon Japanese.* Japan Review, 2007. **19**: 47-75.

167. Turner, H, Hovenkamp, P, and Van Welzen, P, *Biogeography of southeast Asia and the west Pacific.* J Biogeogr, 2001. **28**(2): 217-30.

168. Tougrad, C, *Biogeography and migration routes of large mammal faunas in South-East Asia during the Late Middle Pleistocene: focus on the fossil and extant faunas from Thailand.* Pelaeogeography, Palaeoclimatology, Palaeoecology, 2001. **168**: 337-58.

169. Woodruff, DS and Turner, LM, *The Indochinese-Sundaic zoogeographic transition: a description and analysis of terrestrial mammal species distribution.* Journal of Biogeography, 2009. **36**: 803-21.

170. Woodruff, DS, *Biogeography and conservation in Southeast Asia: how 2.7 million years of repeated environmental fluctuations affect today's patterns and the future of the remaining refugial-phase biodiversity.* Biodivers Conserv, 2010. **19**(4): 919-41.

171. Corlett, RT, *The ecology of tropical East Asia*. 2014, Oxford: Oxford University Press. 291 pp.

172. Soares, P, Trejaut, JA, Loo, JH, Hill, C, Mormina, M, Lee, CL, Chen, YM, Hudjashov, G, Forster, P, Macaulay, V, Bulbeck, D, Oppenheimer, S, Lin, M, and Richards, MB, *Climate change and postglacial human dispersals in Southeast Asia.* Mol Biol Evol, 2008. **25**(6): 1209-18.

173. Blanchon, P and Shaw, J, *Reef drowning during the last deglaciation: evidence for catastrophic sea-level rise and ice-sheet collapse.* Geology, 1995. **23**(1): 4-8.

174. Pelejero, C, Kienast, M, Wang, L, and Grimalt, JO, *The flooding of Sundaland during the last deglaciation: imprints in hemipelagic sediments from the southern South China Sea.* Earth Planet Sc Lett, 1999. **171**(4): 661-71.

175. Bird, MI, Taylor, D, and Hunt, C, *Palaeoenvironments of insular Southeast Asia during the Last Glacial Period: a savanna corridor in Sundaland?* Quat Sci Rev, 2005. **24**(20–21): 2228-42.

176. Solheim, WG, *Archaeology and culture in Southeast Asia: unraveling the Nusantao*. 2006, Quezon City (Philippines): University of Philippines Press. 316 pp.

177. Ooi, KG, *Southeast Asia: a historical encyclopedia, from Angkor Wat to East Timor*. 2004, Santa Barbara: ABC-CLIO. 1791 pp.

178. Bellwood, P, Fox, JJ, and Tryon, D, *The Austronesians: Historical and comparative perspectives*. 2006, Canberra: ANU E Press. 367 pp.

179. Lewis, MP, *Ethnologue: languages of the world*. 16th ed. 2009, Dallas Texas: SIL International. 1248 pp.

180. Zhang, C, *The rise of urbanism in the middle and lower Yangzi River Valley.* Bull Indo Pac Pre Hi, 1997. **16**: 63-67.

181. Chi, Z and Hung, H-C, *The Neolithic of southern China–origin, development, and dispersal.* Asian Perspect, 2008. **47**(2): 299-329.

182. Underhill, AP and Habu, J, *Early communities in East Asia: economic and sociopolitical organization at the local and regional levels*, in *Archaeology of Asia*, Stark, MT, Editor. 2006, Blackwell: Malden. p. 121-48.

183. Fuller, DQ and Qin, L, *Declining oaks, increasing artistry, and cultivating rice: The environmental and social context of the emergence of farming in the Lower Yangtze Region.* Environmental Archaeology, 2010. **15**(2): 139-59.

184. Hsiao-Chun, H, *A sourcing study of Taiwan stone adzes.* Bull Indo Pac Pre Hi, 2004. **24**: 57-70.

185. Bellwood, P, *First farmers: the origins of agricultural societies*. 2004, Oxford: Wiley-Blackwell. 384 pp.

186. Bellwood, P, *Early agriculturalist population diasporas? Farming, languages, and genes.* Annu Rev Anthropol, 2001. **30**: 181-207.

187. Petraglia, M, Clarkson, C, Boivin, N, Haslam, M, Korisettar, R, Chaubey, G, Ditchfield, P, Fuller, D, James, H, Jones, S, Kivisild, T, Koshy, J, Lahr, MM, Metspalu, M, Roberts, R, and Arnold, L, *Population increase and environmental deterioration correspond with microlithic innovations in South Asia ca. 35,000 years ago.* PNAS, 2009. **106**(30): 12261-6.

188. Petraglia, MD, Haslam, M, Fuller, DQ, Boivin, N, and Clarkson, C, *Out of Africa: new hypotheses and evidence for the dispersal of Homo sapiens along the Indian Ocean rim.* Ann Hum Biol, 2010. **37**(3): 288-311.

189. Oppenheimer, S, *Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map.* Philos Trans R Soc Lond B Biol Sci, 2012. **367**(1590): 770-84.

190. Mellars, P, *Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia.* Science, 2006. **313**(5788): 796-800.

191. Mellars, P, Gori, KC, Carr, M, Soares, PA, and Richards, MB, *Genetic and archaeological perspectives on the initial modern human colonization of southern Asia.* PNAS, 2013. **110**(26): 10699-704.

192. Fernandes, V, Alshamali, F, Alves, M, Costa, MD, Pereira, JB, Silva, NM, Cherni, L, Harich, N, Cerny, V, Soares, P, Richards, MB, and Pereira, L, *The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa.* Am J Hum Genet, 2012. **90**(2): 347-55.

193. Haslam, M, Clarkson, C, Petraglia, M, Korisettar, R, Jones, S, Shipton, C, Ditchfield, P, and Ambrose, SH, *The 74 ka Toba super-eruption and southern Indian hominins: archaeology, lithic technology and environments at Jwalapuram Locality 3.* J Archaeol Sci, 2010. **37**(12): 3370-84.

194. Metspalu, M, Kivisild, T, Bandelt, H-J, Richards, M, and Villems, R, *The pioneer settlement of modern humans in Asia*, in *Human Mitochondrial DNA and the Evolution of Homo sapiens*. 2006, Springer: New York. p. 181-99.

195. Bailey, GN, Flemming, NC, King, GC, Lambeck, K, Momber, G, Moran, LJ, Al-Sharekh, A, and Vita-Finzi, C, *Coastlines, submerged landscapes, and human evolution: the Red Sea Basin and the Farasan Islands.* JICA, 2007. **2**(2): 127-60.

196. Barker, G, Barton, H, Beavitt, P, Bird, M, Daly, P, Doherty, C, Gilbertson, D, Hunt, C, Krigbaum, J, and Lewis, H, *Prehistoric foragers and farmers in South-east Asia: Renewed investigations at Niah Cave, Sarawak.* P Prehist Soc, 2002. **68**: 147-64.

197. Bowler, JM, Johnston, H, Olley, JM, Prescott, JR, Roberts, RG, Shawcross, W, and Spooner, NA, *New ages for human occupation and climatic change at Lake Mungo, Australia.* Nature, 2003. **421**(6925): 837-40.

198. O'Connor, S, *New evidence from East Timor contributes to our understanding of earliest modern human colonisation east of the Sunda Shelf.* Antiquity, 2007. **81**(313): 523-35.

199. Demeter, F, Shackelford, L, Westaway, K, Duringer, P, Bacon, A-M, Ponche, J-L, Wu, X, Sayavongkhamdy, T, Zhao, J-X, Barnes, L, Boyon, M, Sichanthongtip, P, Sénégas, F, Karpoff, A-M, Patole-Edoumba, E, Coppens, Y, and Braga, J, *Early modern humans and morphological variation in Southeast Asia: fossil evidence from Tam Pa Ling, Laos.* PLoS ONE, 2015. **10**(4): e0121193.

200. Mijares, AS, Détroit, F, Piper, P, Grün, R, Bellwood, P, Aubert, M, Champion, G, Cuevas, N, De Leon, A, and Dizon, E, *New evidence for a 67,000-year-old human presence at Callao Cave, Luzon, Philippines.* J Hum Evol, 2010. **59**(1): 123-32.

201. Solheim, WG, *II 1984. The Nusantao hypothesis: The origin and spread of Austronesian speakers.* Asian Perspect, 1985. **26**(1): 77-88.

202. Solheim, WG, *Taiwan, coastal south China and northern Viet Nam and the Nusantao maritime trading network.* JEAA, 2000. **2**(1): 273-84.

203. Turner, CG, *Major features of Sundadonty and Sinodonty, including suggestions about East Asian microevolution, population history, and late Pleistocene relationships with Australian aboriginals.* Am J Phys Anthropol, 1990. **82**(3): 295-317.

204. Scott, GR and Turner, CG, *The anthropology of modern human teeth: dental morphology and its variation in recent human populations*. Vol. 20. 2000, Cambridge: Cambridge University Press. 382 pp.

205. Anderson, A, *Crossing the Luzon Strait: archaeological chronology in the Batanes Islands, Philippines and the regional sequence of Neolithic dispersal.* Journal of Austronesian Studies, 2005. **1**(2): 25-44.

206. Higham, C, *Early cultures of mainland Southeast Asia*. 2002, Chicago: Art Media Resources. 375 pp.

207. Bellwood, P and Dizon, E, *Austronesian cultural origins. Out of Taiwan, via the Batanes Islands, and onwards to western Polynesia*, in *Past human migrations in East Asia matching archeology, linguistics and genetics*, Sanchez-Mazas, A, Blench, R, Ross, M, Peiros, P, and Lin, M, Editors. 2008, Routledge: Great Britain. p. 23-39.

208. Diamond, J and Bellwood, P, *Farmers and their languages: the first expansions.* Science, 2003. **300**(5619): 597-603.

209. Blust, R, *The Austronesian languages*. 2009, Canberra: Asia-Pacific Linguistics. 824 pp.

210. Bellwood, P, *A hypothesis for Austronesian origins.* Asian Perspect, 1984: 107-17.

211. Ehret, C, *Bantu expansions: re-envisioning a central problem of early African history.* Int J Afr Hist Stud, 2001. **34**(1): 5-41.

212. Velde, MVd, Nurse, D, Bostoen, K, and Philippson, G, *The Bantu languages*. 2006, London: Routledge. 728 pp.

213. Cavalli-Sforza, LL, Menozzi, P, and Piazza, A, *Demic expansions and human evolution.* Science, 1993. **259**: 639-39.

214. Donohue, M and Denham, T, *Farming and language in Island Southeast Asia.* Curr Anthropol, 2010. **51**(2): 223-56.

215. Gray, RD, Drummond, AJ, and Greenhill, SJ, *Language phylogenies reveal expansion pulses and pauses in Pacific settlement.* Science, 2009. **323**(5913): 479-83.

216. Bellwood, P, *Austronesian prehistory in Southeast Asia: homeland, expansion and transformation*, in *The Austronesians: historical and comparative perspectives*, Bellwood, PS, Fox, JJ, and Tryon, DT, Editors. 1995, ANU E Press: Canberra. p. 103-20.

217. Blust, R, *The prehistory of the Austronesian-speaking peoples: a view from language.* J World Prehist, 1995. **9**(4): 453-510.

218. Ross, M, *The primary subgroups of Austronesian: A reappraisal*, in *Austronesian Historical Linguistics and Culture History: A festschrift for Robert A. Blust*, Adelaar, KA, Pawley, A, and Blust, RA, Editors. 2009, ANU Press: Canberra. p. 295-326.

219. Donohue, M and Grimes, CE, *Yet more on the position of the languages of eastern Indonesia and East Timor.* Ocean Linguist, 2008. **47**(1): 114-58.

220. Dunn, M and Ross, M, *Is Kazukuru really non-Austronesian?* Ocean Linguist, 2007. **46**(1): 210-31.

221. Terrell, JE, Hunt, TL, and Bradshaw, J, *On the location of the proto-Oceanic homeland.* Pacific Studies, 2002. **25**(3): 57-93.

222. Spriggs, M, *The Neolithic and Austronesian expansion within Island Southeast Asia and into the Pacific*, in *From Southeast Asia to the Pacific: Archaeological perspectives on the Austronesian expansion and the Lapita cultural complex*, Chiu, S and Sand, C, Editors. 2007, Academia Sinica: Taipei. p. 104-25.

223. Spriggs, M, *Archaeology and the Austronesian expansion: where are we now.* Antiquity, 2011. **85**(328): 510-28.

224. Rainbird, P, *The archaeology of islands*. 2007, Cambridge: Cambridge University Press. 216 pp.

225. Bellwood, P and Dizon, E, *The Batanes Archaeological Project and the "Out of Taiwan" hypothesis for Austronesian dispersal.* Journal of Austronesian Studies, 2005. **1**(1): 1 - 33.

226. Bedford, S and Sand, C, *Lapita and Western Pacific settlement: progress, prospects and persistent problems*, in *Oceanic explorations: Lapita and western Pacific settlement*, Bedford, S, Sand, C, and Connaughton, SP, Editors. 2007, ANU Press: Canberra. p. 1-15.

227. Jiao, T, *The Neolithic of Southeast China: cultural transformation and regional interaction on the coast*. 2007, New York: Cambria Press. 286 pp.

228. Bellwood, P, *Formosan prehistory and Austronesian dispersal*, in *Austronesian Taiwan: linguistics, history, ethnology, and prehistory*, Bulbeck, D, Editor. 2001, SMC Pub: Taipei. p. 337-65.

229. Flenley, JR, *Palynological evidence for land use changes in South-East Asia.* J Biogeogr, 1988. **15**(1): 185-97.

230. Head, L, *Cultural landscapes and environmental change*. 2000, London: Arnold. 179 pp.

231. Kirch, PV, *Peopling of the Pacific: a holistic anthropological perspective.* Annu Rev Anthropol, 2010. **39**(1): 131-48.

232. Terrell, JE and Welsch, RL, *Lapita and the temporal geography of prehistory.* Antiquity, 1997. **71**(273): 548-72.

233. Bulbeck, D, *An integrated perspective on the Austronesian diaspora: The switch from cereal agriculture to maritime foraging in the colonisation of Island Southeast Asia.* Aust Archaeol, 2008. **67**: 31-52.

234. Meacham, W, *On the improbability of Austronesian origins in South China.* Asian Perspectives, 1984–5. **XXVI**(1): 89–106.

235. Paz, V, *Island southeast Asia - spread or friction zone?*, in *Examining the Farming/Language Dispersal Hypothesis*, Bellwood, P and Renfrew, C, Editors. 2002, MacDonald Institute for Archaeological Research: Cambridge. p. 275-85.

236. Paz, V, *Rock shelters, caves, and archaeobotany in Island Southeast Asia.* Asian Perspect, 2005. **44**(1): 107-18.

237. Barker, G, *The archaeology of foraging and farming at Niah Cave, Sarawak.* Asian Perspect, 2005: 90-106.

238. Hongo, H, Ishiguro, N, Watanobe, T, Shigehara, N, Anezaki, T, The Long, V, Vu Binh, D, Tien, NT, and Nam, NH, *Variation in mitochondrial DNA of Vietnamese pigs: relationships with Asian domestic pigs and Ryukyu wild boars.* Zoolog Sci, 2002. **19**(11): 1329-35.

239. Dobney, K, Cucchi, T, and Larson, G, *The pigs of Island Southeast Asia and the Pacific: new evidence for taxonomic status and human-mediated dispersal.* Asian Perspectives, 2008. **47**: 59-74.

240. Larson, G, Albarella, U, Dobney, K, Rowley-Conwy, P, Schibler, J, Tresset, A, Vigne, J-D, Edwards, CJ, Schlumbaum, A, Dinu, A, Balacsescu, A, Dolman, G, Tagliacozzo, A, Manaseryan, N, Miracle, P, Wijngaarden-Bakker, LV, Masseti, M, Bradley, DG, and Cooper, A, *Ancient DNA, pig domestication, and the spread of the Neolithic into Europe.* Proceedings of the Natural Academy of Science USA, 2007. **104**: 15276-81.

241. O'Connell, JF and Allen, J, *Dating the colonization of Sahul (Pleistocene Australia–New Guinea): a review of recent research.* J Archaeol Sci, 2004. **31**(6): 835-53.

242. Allen, J and O'Connell, JF, *Getting from Sunda to Sahul*, in *Islands of inquiry: colonisation, seafaring and the archaeology of maritime landscapes*, Clark, GR, O'Connor, S, and Leach, BF, Editors. 2008, ANU E Press: Canberra. p. 31.

243. Hill, C, Soares, P, Mormina, M, Macaulay, V, Clarke, D, Blumbach, PB, Vizuete-Forster, M, Forster, P, Bulbeck, D, Oppenheimer, S, and Richards, M, *A mitochondrial stratigraphy for Island Southeast Asia.* Am J Hum Genet, 2007. **80**(1): 29-43.

244. Macaulay, V, Hill, C, Achilli, A, Rengo, C, Clarke, D, Meehan, W, Blackburn, J, Semino, O, Scozzari, R, Cruciani, F, Taha, A, Shaari, NK, Raja, JM, Ismail, P, Zainuddin, Z, Goodwin, W, Bulbeck, D, Bandelt, HJ, Oppenheimer, S, Torroni, A, and Richards, M, *Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes.* Science, 2005. **308**(5724): 1034-6.

245. van Holst Pellekaan, SM, Ingman, M, Roberts-Thomson, J, and Harding, RM, *Mitochondrial genomics identifies major haplogroups in Aboriginal Australians.* Am J Phys Anthropol, 2006. **131**(2): 282-94.

246. van Holst Pellekaan, S, *Genetic evidence for the colonization of Australia.* Quatern Int, 2013. **285**: 44-56.

247. Matisoo-Smith, E, *Ancient DNA and the human settlement of the Pacific: a review.* J Hum Evol, 2015. **79**: 93-104.

248. Soares, P, Rito, T, Trejaut, J, Mormina, M, Hill, C, Tinkler-Hundal, E, Braid, M, Clarke, DJ, Loo, J-H, Thomson, N, Denham, T, Donohue, M, Macaulay, V, Lin, M, Oppenheimer, S, and Richards, MB, *Ancient voyaging and polynesian origins.* Am J Hum Genet, 2011. **88**(2): 239-47.

249. Trejaut, JA, Kivisild, T, Loo, JH, Lee, CL, He, CL, Hsu, CJ, Lee, ZY, and Lin, M, *Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations.* PLoS Biol, 2005. **3**(8): e247.

250. Friedlaender, JS, Friedlaender, FR, Hodgson, JA, Stoltz, M, Koki, G, Horvat, G, Zhadanov, S, Schurr, TG, and Merriwether, DA, *Melanesian mtDNA complexity.* PLoS One, 2007. **2**(2): e248.

251. Barker, G and Richards, MB, *Foraging–farming transitions in Island Southeast Asia.* J Archaeol Method Th, 2012. **20**(2): 256-80.

252. Soares, P, Trejaut, JA, Rito, T, Cavadas, B, Hill, C, Eng, KK, Mormina, M, Brandão, A, Fraser, RM, Wang, T-Y, Loo, J-H, Snell, C, Ko, T-M, Amorim, A, Pala, M, Macaulay, V, Bulbeck, D, Wilson, JF, Gusmão, L, Pereira, L, Oppenheimer, S, Lin, M, and Richards, MB, *Resolving the ancestry of Austronesian-speaking populations.* Hum Genet, 2016. **135**(3): 309-26.

253. Tumonggor, MK, Karafet, TM, Hallmark, B, Lansing, JS, Sudoyo, H, Hammer, MF, and Cox, MP, *The Indonesian archipelago: an ancient genetic highway linking Asia and the Pacific.* J Hum Genet, 2013. **58**(3): 165-73.

254. Ko, AM-S, Chen, C-Y, Fu, Q, Delfin, F, Li, M, Chiu, H-L, Stoneking, M, and Ko, Y-C, *Early Austronesians: into and out of Taiwan.* Am J Hum Genet, 2014. **94**(3): 426-36.

255. Karafet, TM, Hallmark, B, Cox, MP, Sudoyo, H, Downey, S, Lansing, JS, and Hammer, MF, *Major east-west division underlies Y chromosome stratification across Indonesia.* Mol Biol Evol, 2010. **27**(8): 1833-44.

256. Trejaut, JA, Poloni, ES, Yen, J-C, Lai, Y-H, Loo, J-H, Lee, C-L, He, C-L, and Lin, M, *Taiwan Y-chromosomal DNA variation and its relationship with Island Southeast Asia.* BMC Genet, 2014. **15**(1): 77.

257. Stephen Lansing, J, Cox, MP, de Vet, TA, Downey, SS, Hallmark, B, and Sudoyo, H, *An ongoing Austronesian expansion in Island Southeast Asia.* J Anthropol Archaeol, 2011. **30**(3): 262-72.

258. Delfin, F, Myles, S, Choi, Y, Hughes, D, Illek, R, van Oven, M, Pakendorf, B, Kayser, M, and Stoneking, M, *Bridging near and remote Oceania: mtDNA and NRY variation in the Solomon Islands.* Mol Biol Evol, 2012. **29**(2): 545-64.

259. Kayser, M, *Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific.* Mol Biol Evol, 2006. **23**(11): 2234-44.

260. Kayser, M, Choi, Y, van Oven, M, Mona, S, Brauer, S, Trent, RJ, Suarkia, D, Schiefenhovel, W, and Stoneking, M, *The impact of the Austronesian expansion: evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of Melanesia.* Mol Biol Evol, 2008. **25**(7): 1362-74.

261. Hage, P and Marck, J, *Matrilineality and the Melanesian origin of Polynesian Y chromosomes.* Curr Anthropol, 2003. **44**(S5): S121-S27.

262. Xu, S, *Human population admixture in Asia.* Genomics Inform, 2012. **10**(3): 133-44.

263. Xu, S, Pugach, I, Stoneking, M, Kayser, M, and Jin, L, *Genetic dating indicates that the Asian-Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion.* PNAS, 2012. **109**(12): 4574-79.

264. Jinam, TA, Hong, LC, Phipps, ME, Stoneking, M, Ameen, M, Edo, J, and Saitou, N, *Evolutionary history of continental Southeast Asians: "Early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data.* Mol Biol Evol, 2012. **29**(11): 3513-27.

265. Lipson, M, Loh, P-R, Patterson, N, Moorjani, P, Ko, Y-C, Stoneking, M, Berger, B, and Reich, D, *Reconstructing Austronesian population history in Island Southeast Asia.* Nat Commun, 2014. **5**: 4689.

266. Mirabal, S, Cadenas, AM, Garcia-Bertrand, R, and Herrera, RJ, *Ascertaining the role of Taiwan as a source for the Austronesian expansion.* Am J Phys Anthropol, 2013. **150**(4): 551-64.

267. Pandey, JP, *Genomewide association studies and assessment of risk of disease.* N Engl J Med, 2010. **363**(21): 2076-7; author reply 77.

268. Corona, E, Chen, R, Sikora, M, Morgan, AA, Patel, CJ, Ramesh, A, Bustamante, CD, and Butte, AJ, *Analysis of the Genetic Basis of Disease in the Context of*

Worldwide Human Relationships and Migration. PLoS Genet, 2013. **9**(5): e1003447.

269. McPherson, R, *From genome-wide association studies to functional genomics: new insights into cardiovascular disease.* Can J Cardiol, 2012. **29**(1): 23-29.

270. Myles, S, Davison, D, Barrett, J, Stoneking, M, and Timpson, N, *Worldwide population differentiation at disease-associated SNPs.* BMC Med Genomics, 2008. **1**: 22.

271. Rosenberg, NA, Huang, L, Jewett, EM, Szpiech, ZA, Jankovic, I, and Boehnke, M, *Genome-wide association studies in diverse populations.* Nat Rev Genet, 2010. **11**(5): 356-66.

272. Hancock, AM, Witonsky, DB, Gordon, AS, Eshel, G, Pritchard, JK, Coop, G, and Di Rienzo, A, *Adaptations to climate in candidate genes for common metabolic disorders.* PLoS Genet, 2008. **4**(2): e32.

273. Young, JH, Chang, YP, Kim, JD, Chretien, JP, Klag, MJ, Levine, MA, Ruff, CB, Wang, NY, and Chakravarti, A, *Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion.* PLoS Genet, 2005. **1**(6): e82.

274. Dizon, E, xe, troit, F, xe, mah, F, xe, ois, Falgu, xe, res, C, Hameau, S, xe, bastien, Ronquillo, W, and Cabanis, E, *Notes on the morphology and age of the Tabon Cave fossil Homo sapiens.* Curr Anthropol, 2002. **43**(4): 660-66.

275. Détroit, F, Dizon, E, Falguères, C, Hameau, S, Ronquillo, W, and Sémah, F, *Upper Pleistocene Homo sapiens from the Tabon cave (Palawan, The Philippines): description and dating of new discoveries.* Comptes Rendus Palevol, 2004. **3**(8): 705-12.

276. HUGO Pan-Asian SNP Consortium, *Mapping human genetic diversity in Asia.* Science, 2009. **326**(5959): 1541-45.

277. Tabbada, KA, Trejaut, J, Loo, JH, Chen, YM, Lin, M, Mirazon-Lahr, M, Kivisild, T, and De Ungria, MC, *Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia?* Mol Biol Evol, 2010. **27**(1): 21-31.

278. Pawley, A, *The Austronesian dispersal: languages, technologies and people*, in *Examining the farming/language dispersal hypothesis*, Bellwood, P and Renfrew, C, Editors. 2002, McDonald Institute for Archaeological Research: Cambridge. p. 251–73.

279. Bellwood, P, *The prehistory of the Indo-Pacific archipelago*. 1997: Honolulu: University of Hawaii Press. 384 pp.

280. Forestier, H, Simanjuntak, T, Guillaud, D, Driwantoro, D, Wiradnyana, K, Siregar, D, and Awe, R, *The site of Togi Ndrawa, Island of Nias, North Sumatra: the first record of a Hoabinhian cave occupation in Indonesia.* Comptes Rendus Palevol, 2005. **4**(8): 727-33.

281. Bulbeck, FD and Mercader, J, *Hunter-gatherer occupation of the Malay Peninsula from the Ice Age to the Iron Age*, in *Under the canopy: the archaeology of tropical rain forests*, Mercader, J, Editor. 2002, Rutgers University Press: Piscataway. p. 119-60.

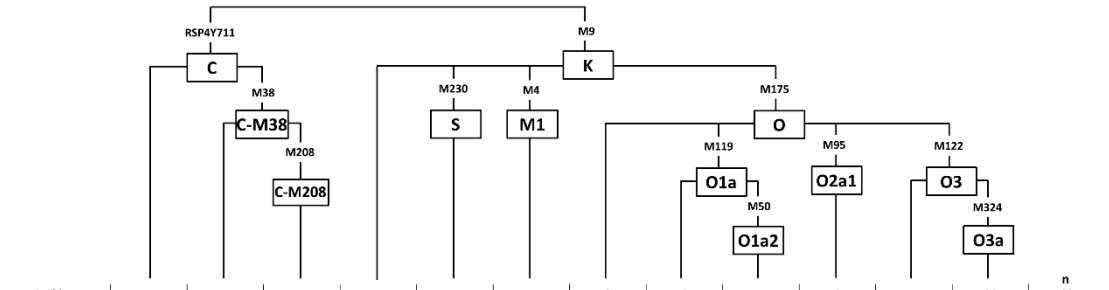
282. Mijares, ASB, *Unearthing prehistory: the archaeology of Northeastern Luzon, Philippine Islands*. Vol. 1613. 2007: British Archaeological Reports Limited. 139 pp.

283. Rootsi, S, Behar, DM, Jarve, M, Lin, AA, Myres, NM, Passarelli, B, Poznik, GD, Tzur, S, Sahakyan, H, Pathak, AK, Rosset, S, Metspalu, M, Grugni, V, Semino, O, Metspalu, E, Bustamante, CD, Skorecki, K, Villems, R, Kivisild, T, and Underhill, PA, *Phylogenetic applications of whole Y-chromosome sequences and the Near Eastern origin of Ashkenazi Levites.* Nat Commun, 2013. **4**: 2928.

APPENDICES

APPENDIX A – SUPPLEMENTARY INFORMATION OF PAPER I

Resolving the ancestry of Austronesian-speaking populations



Location/Group													n
S. China							2	4		1		16	23
Ami								9	14	2	2	22	49
Atayal								27	8				35
Bunun								4	17	4			25
Paiwan							1	44	19			1	65
Yami								7		3			10
Philippines								7	1		2	13	23
Pekanbaru	1							10	3	8	2	7	31
Kota Kinabalu	3	3						11	7	7	3	10	44
Banjarmasin	8			3				1	2	7		4	25
Palu		2			3	1		4		5		13	28
Toraja	1	2		1	5	3		4	16	8		8	48
Mataran	3	2		2	2	2		3	4	16		4	38
Alor		21		4		11		1	2			7	46
Madang		3	9	22	10	28				1			73
Lihir		10		21	2	2			2				41
Kavieng		4		7	1	13			1			4	26
Vanuatu		9	5	33	7	10		1	2			3	70
Fiji	2	4	8	6		3						3	26
Atiu			33	3								1	37
French Polynesia			18	4				1				11	34
Total													797

Figure S1. Y-chromosome tree of the SNPs analysed. The embedded table indicates the distribution of the haplogroups across the sampled area.

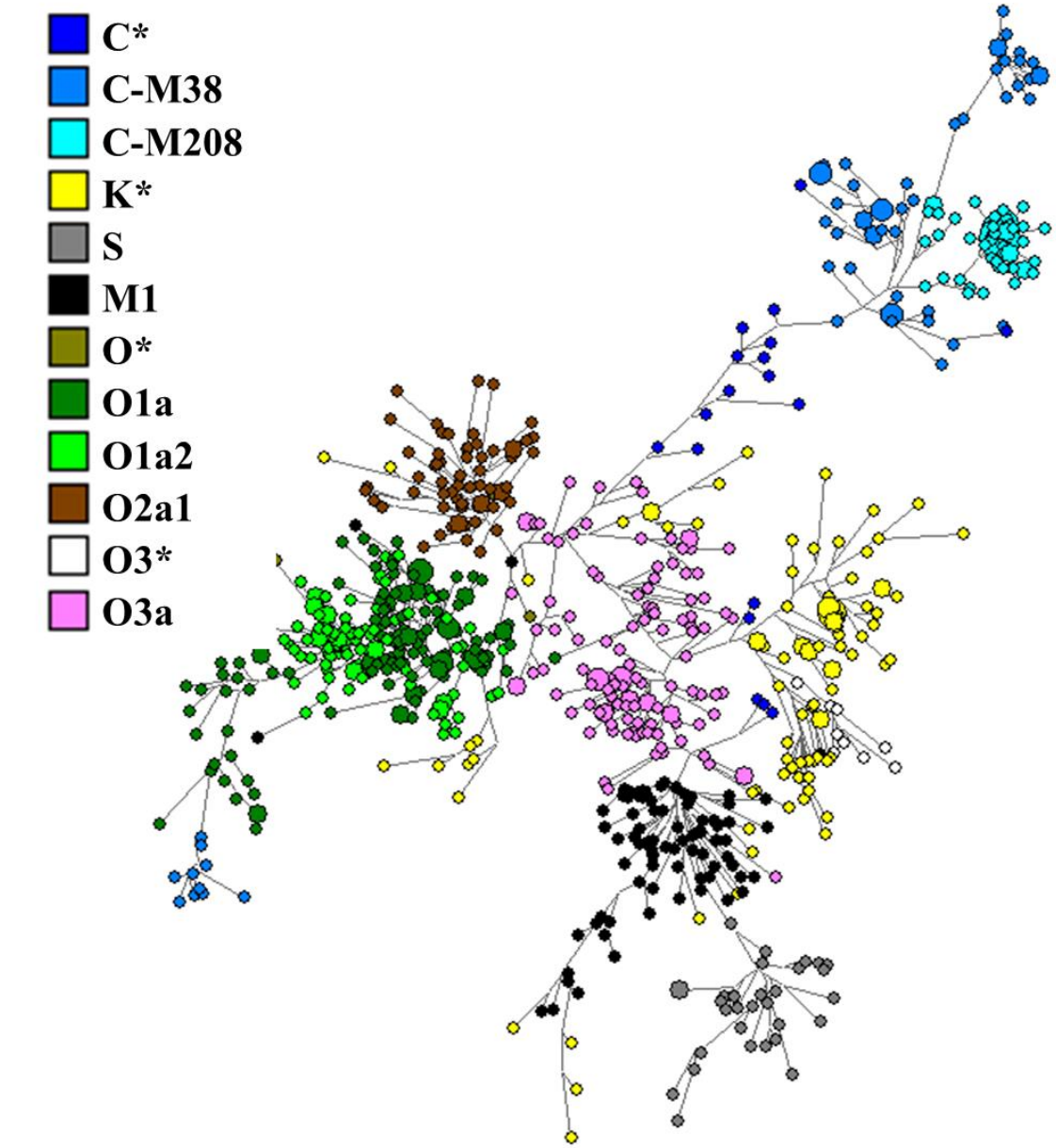


Figure S2. Overall Y-chromosome STR network, calculated using the median-joining algorithm. SNPs were not included in the phylogenetic reconstruction and the samples were labelled according to their SNP lineage after the network construction, to test the robustness of the phylogeny.

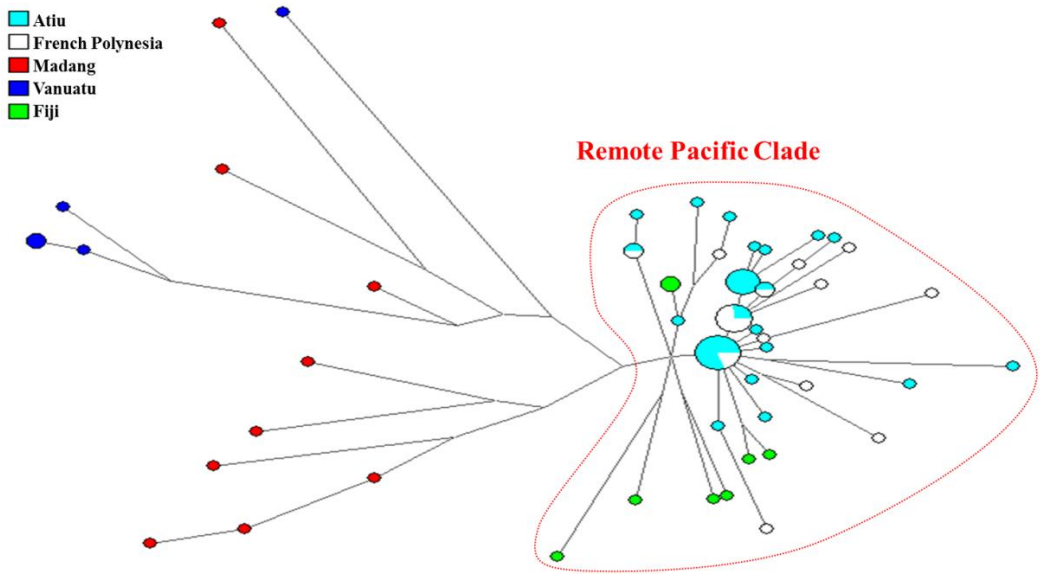


Figure S3. STR network of haplogroup C-M208, indicating the subclade that is exclusive to the Remote Pacific

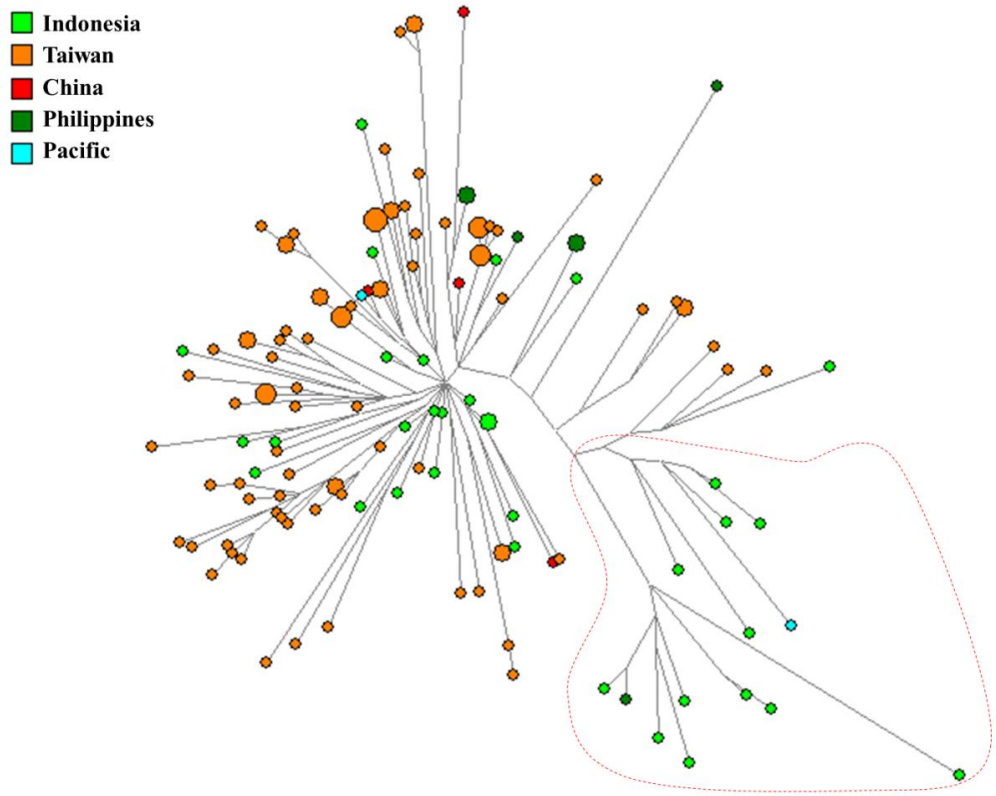


Figure S4. STR network of haplogroup O1*. A subclade displaying a deeper ancestry in ISEA than the remainder of the haplogroup is indicated as indicated by the founder analysis.

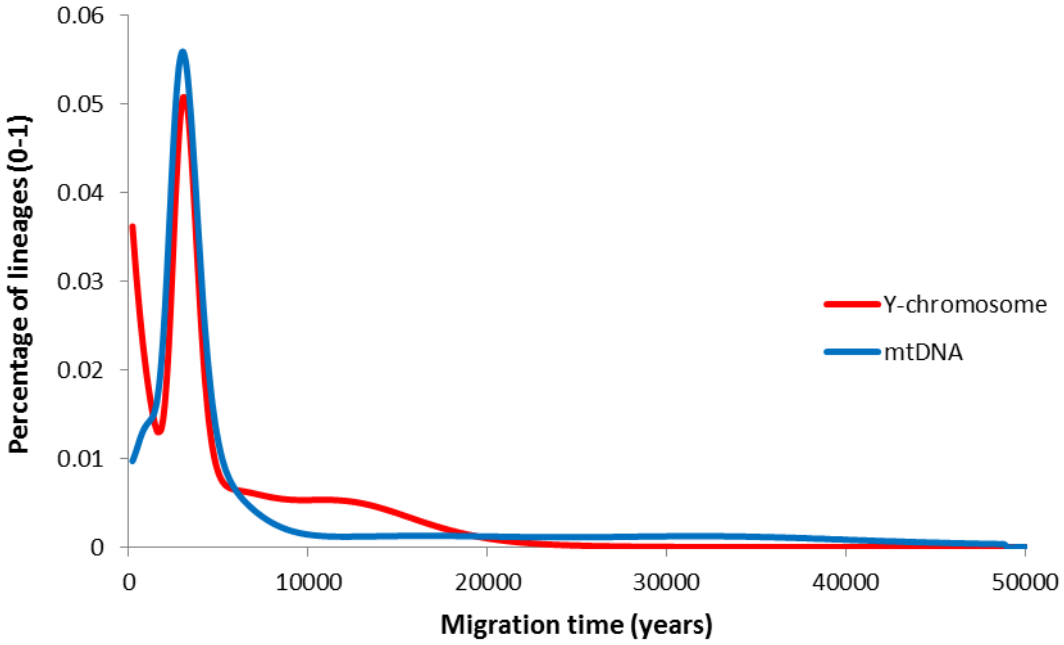


Figure S5. Scan of migration time from ISEA/Near Oceania into Remote Oceania using both Y-chromosome and mtDNA variation

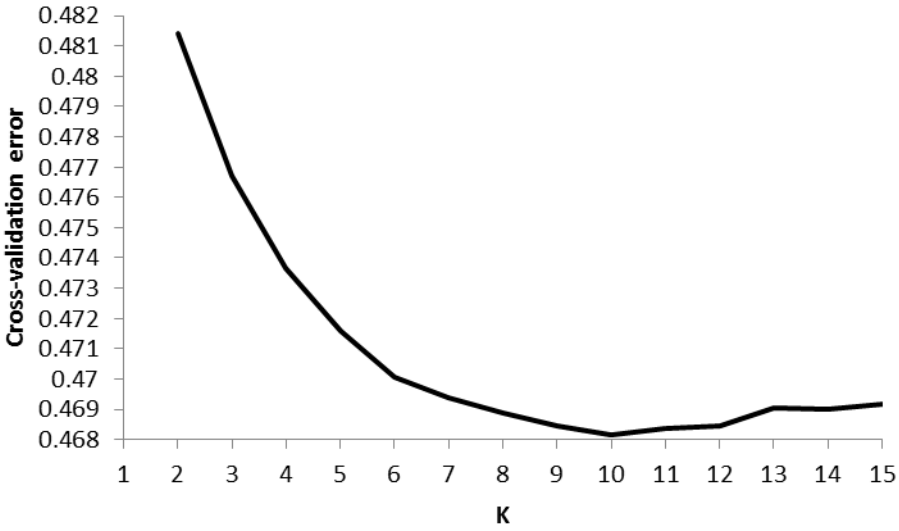


Figure S6. Plot of cross-validation errors across different analyses of ADMIXTURE, against different numbers of ancestral populations (K)

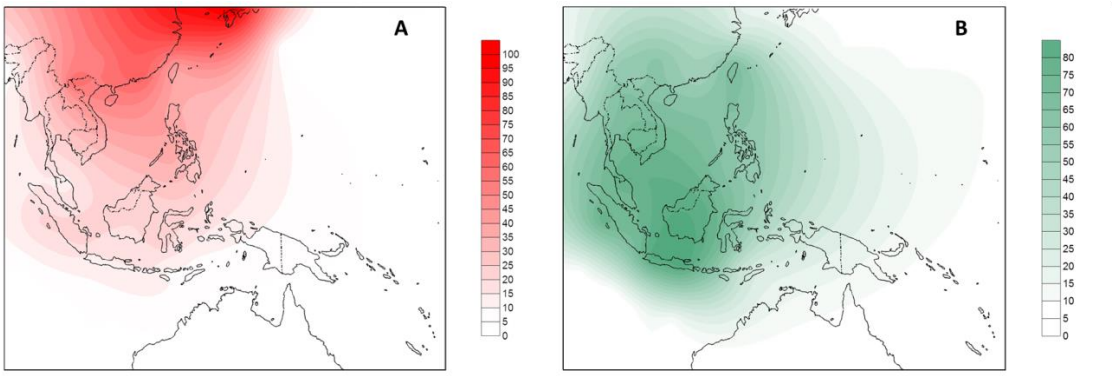


Figure S7. Frequency distribution maps of the two East Asian components obtained on the A*DMIXTURE* analysis when five ancestral populations were considered. The outline map was obtained from www.outline-world-map.com.

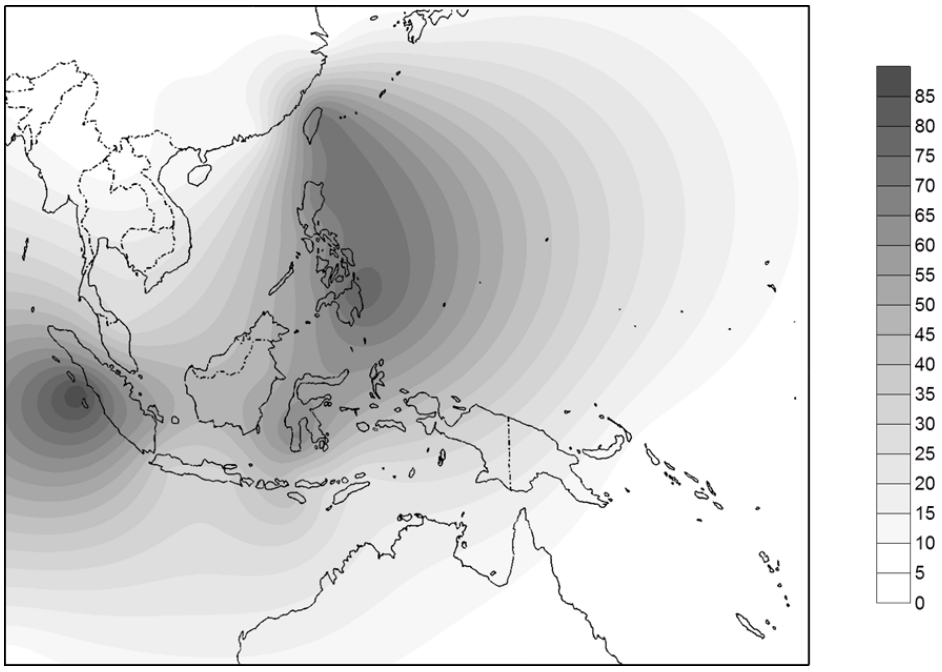


Figure S8. Frequency distribution map of an Island Southeast Asian/Taiwanese component obtained on the ADMIXTURE analysis when 10 ancestral populations were considered. The outline map was obtained from www.outline-world-map.com.

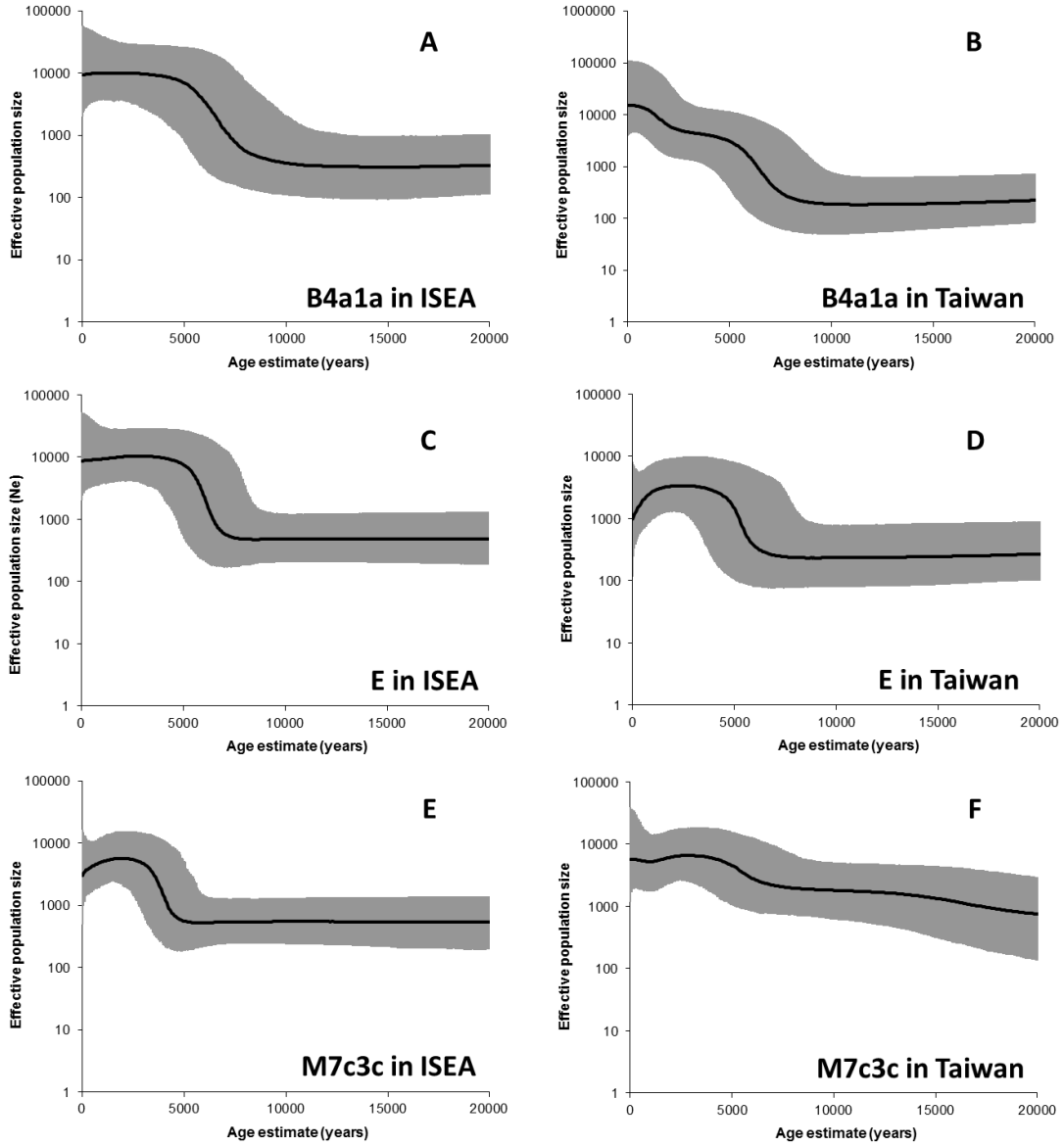


Figure S9. Bayesian skyline plots (BSPs) for haplogroups B4a1a, E and M7c3c in ISEA and Taiwan

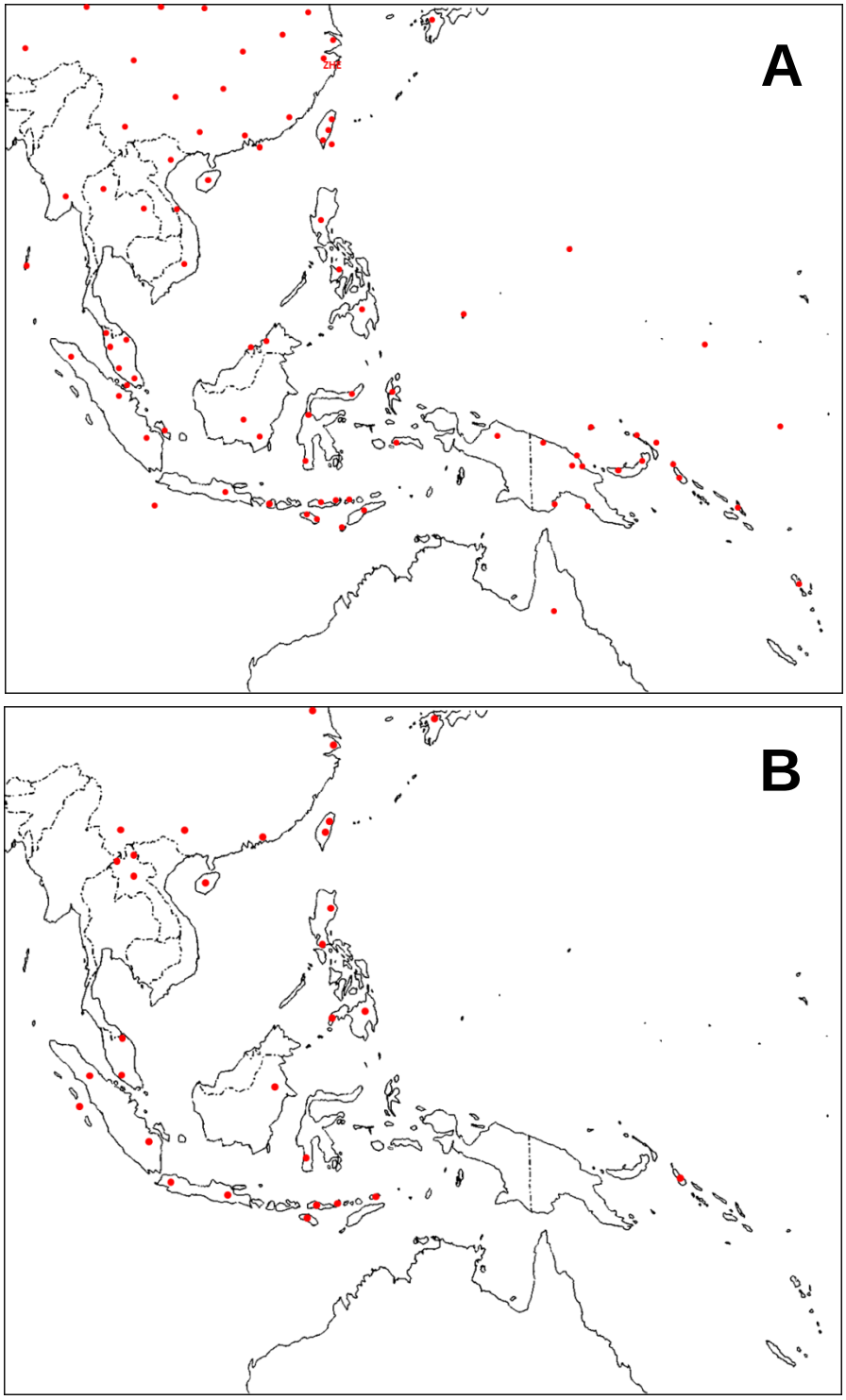


Figure S10. Data points used in the Surfer software for obtaining the frequency distribution of mtDNA clades (A) and autosomal components (B). The outline map was obtained from www.outline-world-map.com.

Table S1. Source and sink mtDNA HVS-I datasets employed in the mtDNA founder analysis into ISEA.

Due to the large size of the table S1, the data is provided in digital format (Supplementary Tables – Paper I)

Table S2. Additional data compiled and eventually used to refine the topology of the HVS-I networks but not employed either as source or sink population in any analysis.

Region	**Sub-region/group**	***n***	**Reference**
Andaman islands	Great Andamanese	20	[1]
	Jarawa	4	[1]
	Onge	63	[1]
Australia	Unknown	54	[2]
	Darling River, West	63	[3]
	Kimberley of Western Australia	2	[4]
	western desert of Western Australia	2	[4]
	Yuendumu, Central Australia	51	[3]
	northwestern Australia	32	[5]
Singapore	-	55	Unp.
Malaysia	Johor	71	[6, 7]
	Kedah/Perlis/Penang	52	[6, 7]
	Perak	67	[6, 7]
	Kelantan/Terengganu	106	[6, 7]
	Selangor/Wilayah/Negeri/Melaka	223	[6-8]
Orang Asli	Malaysia	288	[9]; Unp
	Sakai	20	[10]
Christmas Islands	Christmas Islands	70	[11]
Micronesia	Guam	40	Unpublished
	Nauru	34	Unpublished
	Kiribati	14	Unpublished

Table S3. Source and sink mtDNA HVS-I datasets employed in the mtDNA founder analysis into Remote Oceania. Both source and sink populations in Table S1 are included in the source for this analysis.

Region	Sub-region/group	*n*	Reference
Source			
Karkar Islands		47	[12]
New Guinea	Simbu/Western Highlands	16	Unpublished
	Bundi	58	[5]; Unp
	Irian Jaya	178	Unpublished
	Southern Highlands	17	Unpublished
	Sepik Province	219	[13]
	Port Moresby	117	Unpublished
	Madang	163	Unpublished
	Undetermined	78	[2]; Healy and Hunley (genbank direct submission); Unpublished
Bismarck Archipelago	Balopa	59	[14]
	East New Britain	222	[15]
	West New Britain	353	[15]
	Lavongai	18	[15]
	Kavieng	83	Unpublished
	Lihir	94	Unpublished
	New Ireland Papua	62	[15]
	North New Ireland Astronesian	98	[15]
	Mussau	16	[15]
Bougainville	South	109	[15]
	North	91	[15]
	Central - Aita	33	[15]
	-	22	Healy and Hunley (genbank direct submission); Unpublished
Solomon Islands	Malaita	237	[15]
	-	21	Unp
Sink			
Vanuatu	-	130	[16]; Unpublished
New Zealand	-	13	Pierson and Fris (genbank direct submission); Unpublished
Cook Islands		27	Pierson and Fris (genbank direct submission); Unpublished
Fiji	-	1	Pierson and Fris, 2006
Tonga	-	51	[14]; Pierson and Fris (genbank direct submission);
Samoa	-	39	[2]; Pierson and Fris (genbank direct submission);
French Polynesia	Mangareva	17	[17], Unpublished

Table S4. Primers used in the typing of ten Y-STRs, including the fluorescence label for each forward primer (FAM, TET, HEX). References are provided when the primers were taken from the literature.

STR	Primer Forward	Reverse Primer
DYS460	FAM-AGCAAGCACAAGAATACCAGAG [18]	TCTATCCTCTGCCTATCATTTATTA [19]
DYS461	FAM-AGGCAGAGGATAGATGATATGGAT [19]	TGATGCTGTGTCACTATATTTCTG [18]
DYS438	FAM-TGGGGAATAGTTGAACGGTAA [20]	GTGGCAGACGCCTATAATCC [20]
DYS448	FAM- TGTCAAAGAGCTTCAATGGAGA (*)	TCTTCCTTAACGTGAATTTCCTC (*)
DYS425	TET- TGGAGAGAAGAAGAGAGAAAT (*)	AGTAATTCTGGAGGTAAAATGG (*)
DYS458	TET-GCAACAGGAATGAAACTCCAAT (*)	GTTCTGGCATTACAAGCATGAG (*)
DYS437	TET-GACTATGGGCGTGAGTGCAT [20]	AGACCCTGTCATTCACAGATGA [20]
DYS439	TET-TCCTGAATGGTACTTCCTAGGTTT [20]	GCCTGGCTTGGAATTCTTTT [20]
GATA-H4	TET-GTTATGCTGAGGAGAATTTCCAA [18]	CCTCTGATGGTGAAGTAATGGAATTAGA [19]
DYS388	HEX – GTGAGTTAGCCGTTTAGCGA (*)	CAGATCGCAACCACTGCG (*)
GATA-A10	HEX-CCTGCCATCTCTATTTATCTTGC (*)	TGGAGATAGTGGGTGGATTGA(*)
DYS635	HEX-AGTGTCTCACTTCAAGCACCAAGCAC [19]	GCAGCAAAATTCACAGTTGGAAAAATGT [19]

(*) Newly designed primer

Table S5. Primers and restriction enzymes used in the typing of three Y-chromosome SNPs.

SNP	Primer Forward	Reverse Primer	Restriction Enzyme
M208	GCAACGATTTATCAGCTTTCA	GCAGGAAAAGCCTGTTTGTT	*TaqI*
M230	AATGTCACATTTAGTCTTAACCCAT	ACATTATTAGTATGTAAATCTTCATTGC	*Tsp5091*
M324	TGATAGAAGGCAAGAGGGAGT	AACAAATTGATTTCCAGGGATA	*MnlI*

Table S6. Samples used in the ADMIXTURE analysis.

Population	code	n	Ethnicity	Locale
Yoruba	YRI	60	Yoruba	Nigeria
India	IN-WI	25	Caucasoids	Rajasthan, India
	IN-WL	14	Caucasoids	Maharashtra, India
Japanese	JP-ML	71	Japanese	Tokyo, Japan
	JPT	44	Japanese	Tokyo, Japan
Koreans	KR-KR	90	Koreans	Gyunggi-province, Korea
Han	CN-SH	21	Han	Shanghai, China
Han	CHB	45	Han	Beijing, China
Chinese in Taiwan	TW-HA	48	Chinese	Taipei, Taiwan
	TW-HB	32	Chinese	Taipei, Taiwan
Han	CN-GA	30	Han	Guangzhou, China
Zhuang	CN-CC	26	Zhuang	Guangxi, China
Jiamao	CN-JI	31	Jiamao	Hainan, China
Wa	CN-WA	29	Wa	Yunnan, China
Wa	CN-WA	27	Wa	Yunnan, China
Jinuo	CN-JN	29	Jinuo	Yunnan, China
Yao	TH-YA	19	Yao	Chiang Rai province, Thailand
				Phayao province, Thailand
				Nan province, Thailand
Paluang	TH-PL	18	Paluang	Chiang Mai province, Thailand
Karen	TH-KA	20	Karen	Mae Hong Son province, Thailand
				Chiang Mai province, Thailand
Lawa	TH-LW	19	Lawa	Mae Hong Son province, Thailand
Tai	TH-TU	20	Tai Yuan	Lamphun province, Thailand
				Chiang Mai province, Thailand
				Saraburi province, Thailand
	TH-TY	18	Tai Yong	Lamphun province, Thailand
	TH-TL	20	Tai Lue	Nan province, Thailand
				Chiang Mai province, Thailand
	TH-TK	18	Tai Khuen	Chiang Mai province, Thailand
Ami	AX-AM	10	Ami	Taiwan
Atayal	AX-AT	10	Atayal	Taiwan
Filipino	PI-UB	20	Filipino	Isabela Province, The Philippines
Filipino	PI-UN	19	Filipino	Metro Manila, The Philippines
Minanubu	PI-MA	18	Minanubu	Loreto, Agusan del Sur, The Philippines
Filipino	PI-UI	20	Filipino	Zamboango, The Philippines
Proto-Malay	MY-TM	49	Proto-Malay	Jelebu District, Negri Sembilan, Malaysia
			Proto-Malay	Kuala Pilah District, Negri Sembilan, Malaysia
Malay	MY-KN	18	Malay	Jeli (Dabung), Machang, Kelantan, Malaysia
Malay	MY-MN	20	Malay	Lenggeng, Negeri Sembilan, Malaysia
Dayak	ID-DY	12	Dayak	East Kalimantan, Indonesia
Batak	ID-TB	20	Batak Toba	Balige, Sumatra, Indonesia
	ID-KR	17	Batak Karo	Karo, North Sumatra, Indonesia
Malay	ID-ML	12	Malay	Pelembang, South Sumatra, Indonesia
Mentawai	ID-MT	15	Mentawai	Mentawai Island, Indonesia
Sunda	ID-SU	25	Sunda	Jakarta, Java, Indonesia
Javanese	ID-JV	19	Javanese	Java, Indonesia
	ID-JA	34	Javanese	Jakarta, Java, Indonesia
Toraja	ID-TR	20	Toraja	Tana Toraja, Sulawesi, Indonesia
Kambera	ID-SB	20	Kambera	Sumba Timur, Indonesia
Manggarai	ID-SO	19	Manggarai	Ngada, Flores, Indonesia
	ID-RA	17	Manggarai	Rampasasa, Manggarai, Indonesia
Lamaholot	ID-LA	20	Lamaholot	Larantuka, East Flores, Indonesia
Alorese	ID-AL	19	Alorese	Alor Island, Indonesia
Lembata	ID-LE	19	Lembata	Lembata, East Flores, Indonesia
Melanesians	AX-ME	5	Melanesians	Indo-Pacific

Table S7. M7 sequences used in the phylogenetic reconstruction.

Due to the large size of the table S7, the data is provided in digital format (Supplementary Tables – Paper I):

Table S8. M9/E sequences used in the phylogenetic reconstruction.

Due to the large size of the table S8, the data is provided in digital format (Supplementary Tables – Paper I)

Table S9. B4a1a sequences used in the phylogenetic reconstruction.

Due to the large size of the table S9, the data is provided in digital format (Supplementary Tables – Paper I)

Table S10. Sequences used in the ancient DNA fossil calibration with BEAST.

Due to the large size of the table S10, the data is provided in digital format (Supplementary Tables – Paper I)

Table S11. Increment periods, peak of increment and ratio of increment in the Bayesian skyline plots (BSPs) of mtDNA haplogroups B4a1a, E and M7c3c in ISEA and Taiwan.

Clade	Location	Increment period	Ratio of increment	Peak
B4a1a	*ISEA*	3.5-10.2 ka	21x	6.7 ka
	Taiwan	0.4-9.3 ka	85x	6.7 ka; 1.5 ka
E	*ISEA*	3.8-7.7 ka	11.5x	6.1 ka
	Taiwan	3.1-7.4 ka	8.9x	5.2 ka
M7c3c	*ISEA*	2.2-5.2 ka	7.6x	4 ka
	Taiwan	3.6-7.6 ka	2.9x	5.2 ka

References

1. Thangaraj, K., et al., *Genetic affinities of the Andaman Islanders, a vanishing human population.* Current Biology, 2003. **13**(2): p. 86-93.
2. Redd, A.J. and M. Stoneking, *Peopling of Sahul: mtDNA variation in Aboriginal Australian and Papua New Guinean populations.* American Journal of Human Genetics, 1999. **65**(3): p. 808-828.
3. Van Holst Pellekaan, S.M., et al., *Mitochondrial control-region sequence variation in aboriginal Australians.* American Journal of Human Genetics, 1998. **62**(2): p. 435-449.
4. Betty, D.J., et al., *Multiple independent origins of the COII/tRNA(Lys) intergenic 9-bp mtDNA deletion in aboriginal Australians [4].* American Journal of Human Genetics, 1996. **58**(2): p. 428-433.
5. Hudjashov, G., et al., *Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis.* Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(21): p. 8726-8730.
6. Zainuddin, Z. and W. Goodwin, *Mitochondrial DNA profiling of modern Malay and Orang Asli populations in peninsular Malaysia.* International Congress Series, 2004. **1261**(0): p. 428-430.
7. Nur Haslindawaty, A.R., et al., *Sequence polymorphisms of mtDNA HV1, HV2, and HV3 regions in the Malay population of Peninsular Malaysia.* International Journal of Legal Medicine, 2010. **124**(5): p. 415-426.
8. Maruyama, S., et al., *MtDNA control region sequence polymorphisms and phylogenetic analysis of Malay population living in or around Kuala Lumpur in Malaysia.* International Journal of Legal Medicine, 2010. **124**(2): p. 165-170.
9. Macaulay, V., et al., *Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes.* Science, 2005. **308**(5724): p. 1034-6.
10. Fucharoen, G., S. Fucharoen, and S. Horai, *Mitochondrial DNA polymorphisms in Thailand.* Journal of Human Genetics, 2001. **46**(3): p. 115-125.
11. Wise, C.A., et al., *Y-chromosome and mitochondrial DNA studies on the population structure of the Christmas Island Community.* American Journal of Physical Anthropology, 2005. **128**(3): p. 670-677.
12. Ricaut, F.X., et al., *Mitochondrial DNA variation in Karkar Islanders.* Annals of Human Genetics, 2008. **72**(3): p. 349-367.
13. Vilar, M.G., et al., *Reconstructing the origin of the Lapita Cultural Complex: mtDNA analyses of East Sepik Province, PNG.* Journal of Human Genetics, 2008. **53**(8): p. 698-708.
14. Ohashi, J., et al., *Brief communication: Mitochondrial DNA variation suggests extensive gene flow from Polynesian ancestors to indigenous Melanesiens in the northwestern Bismarck Archipelago.* American Journal of Physical Anthropology, 2006. **130**(4): p. 551-556.
15. Friedlaender, J.S., et al., *Melanesian mtDNA complexity.* PLoS ONE, 2007. **2**(2): p. e248.
16. Hagelberg, E., et al., *Evidence for mitochondrial DNA recombination in a human population of island Melanesia.* Proceedings of the Royal Society B: Biological Sciences, 1999. **266**(1418): p. 485-492.

17. Deguilloux, M.F., et al., *Human ancient and extant mtDNA from the Gambier Islands (French polynesia): Evidence for an early Melanesian maternal contribution and new perspectives into the settlement of Easternmost Polynesia.* American Journal of Physical Anthropology, 2011. **144**(2): p. 248-257.

18. Sanchez-Diz, P., et al., *Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study.* International Journal of Legal Medicine, 2008. **122**(6): p. 529-533.

19. White, P.S., et al., *New, male-specific microsatellite markers from the human Y chromosome.* Genomics, 1999. **57**(3): p. 433-437.

20. Ayub, Q., et al., *Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information.* Nucleic acids research, 2000. **28**(2): p. e8.

APPENDIX B – SUPPLEMENTARY INFORMATION OF PAPER II

Quantifying the legacy of the Chinese Neolithic on the maternal genetic heritage of Taiwan and Island Southeast Asia

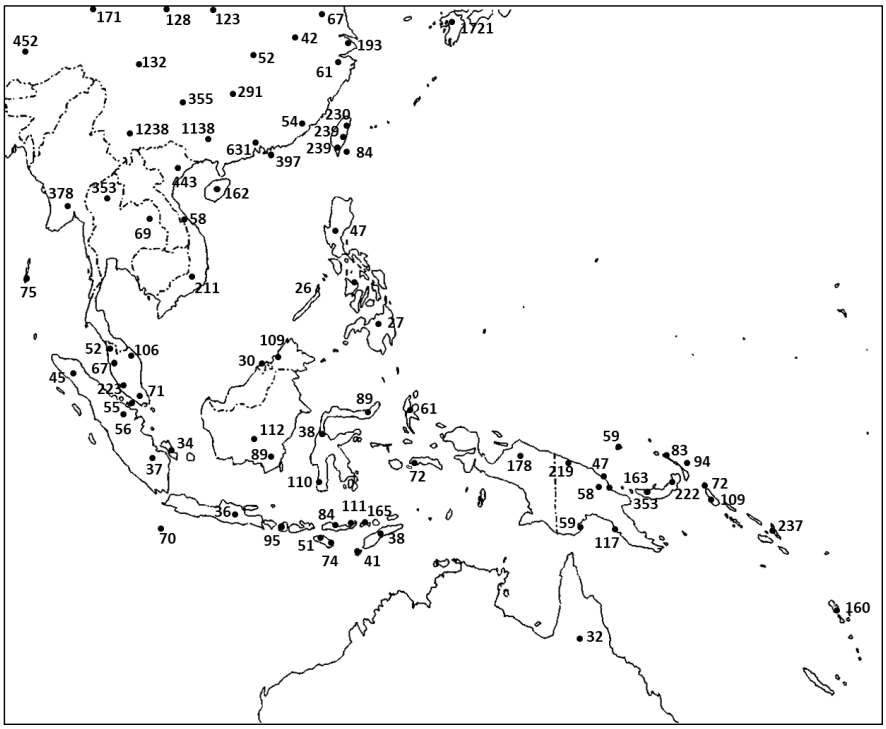


Figure S1. Map showing the geographic distribution and the sample sizes for the dataset used in the Surfer analyses.

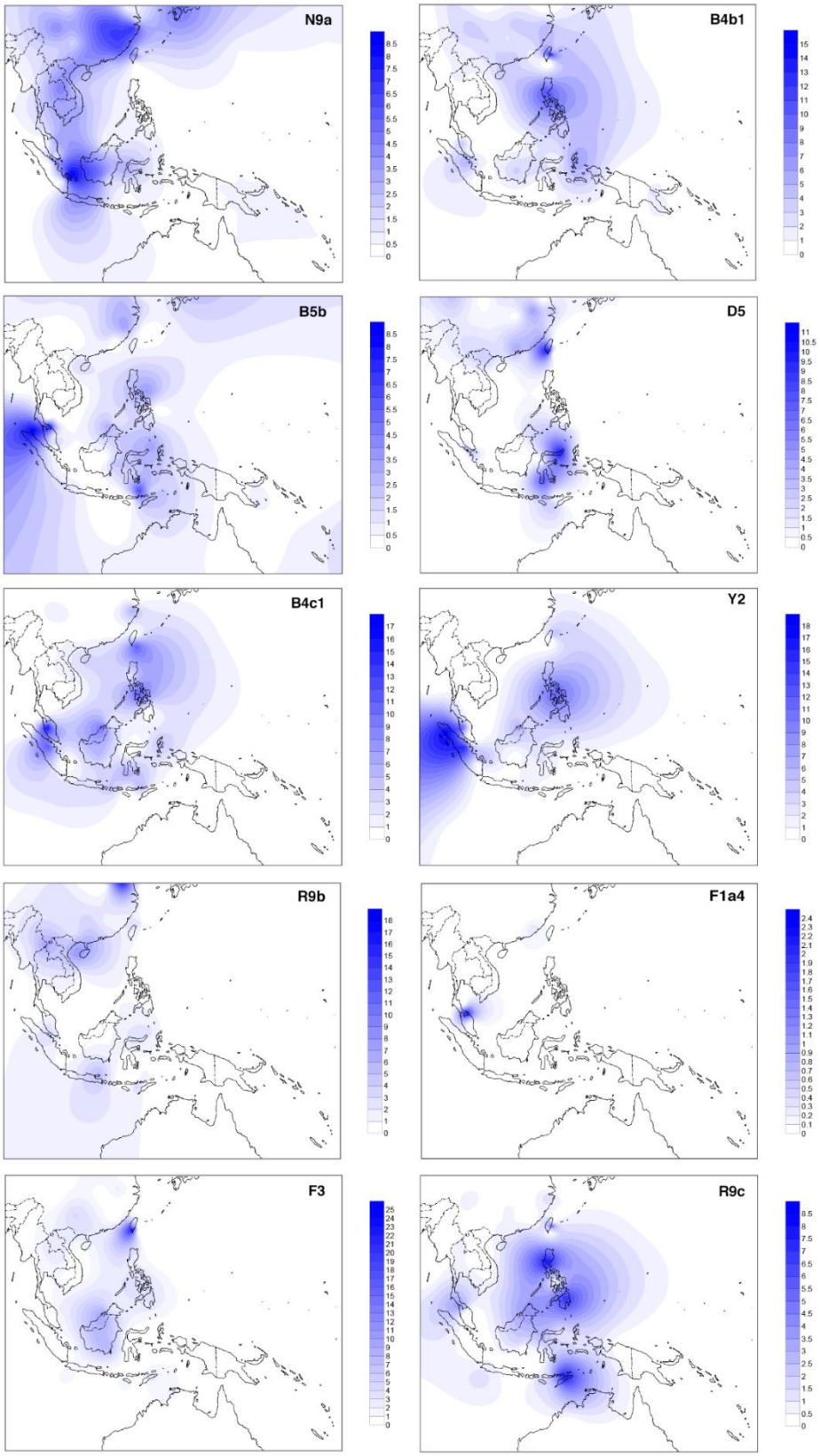


Figure S2. Frequency distribution maps for mtDNA haplogroups examined in this study based on HVS-I data. Map created using Surfer.

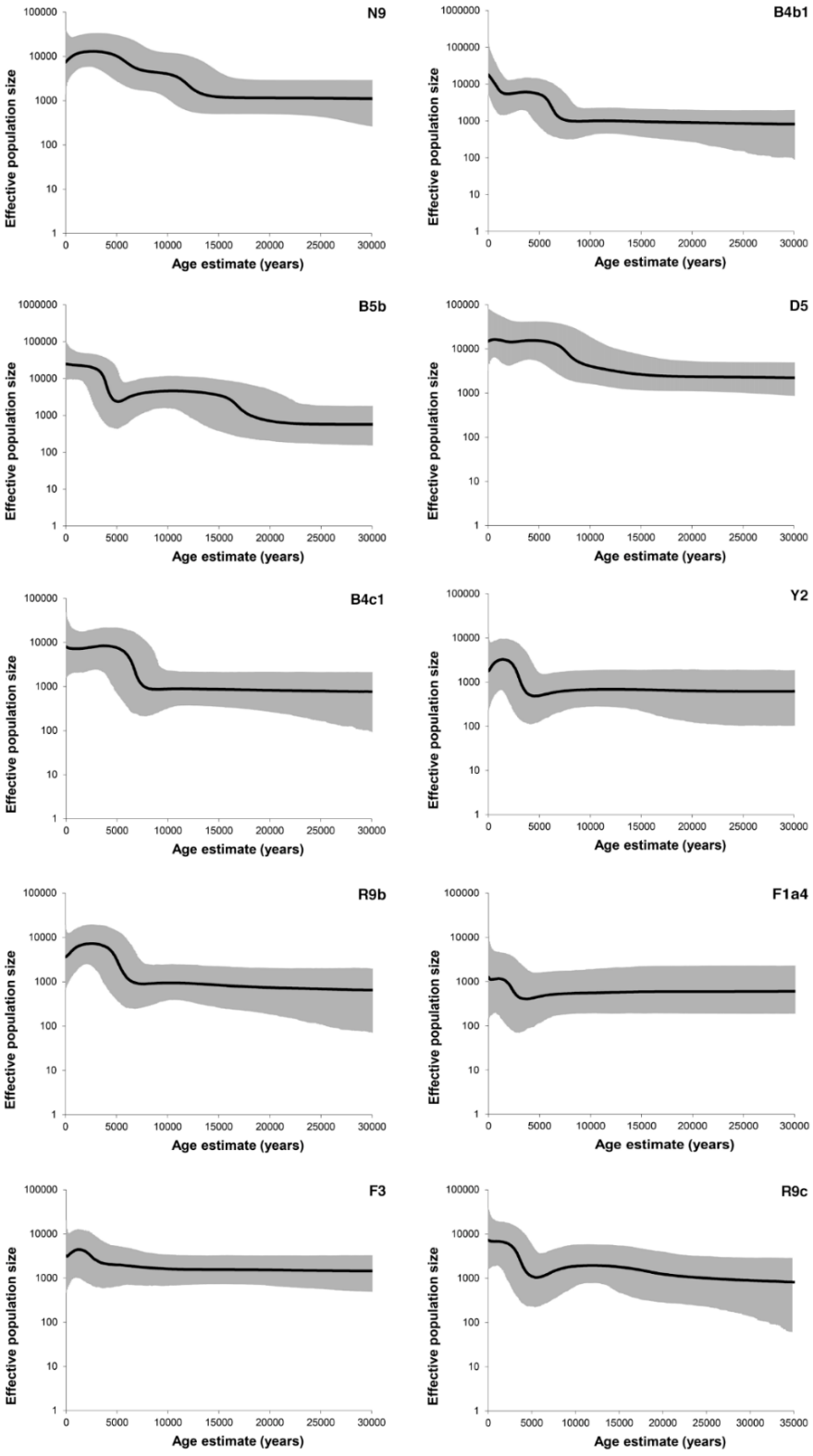


Figure S3. Bayesian skyline plots for mtDNA haplogroups examined in this study, assuming a generation of 25 years. The black lines represent the posterior median of the effective population size through time, and the grey regions represent the 95% confidence interval.

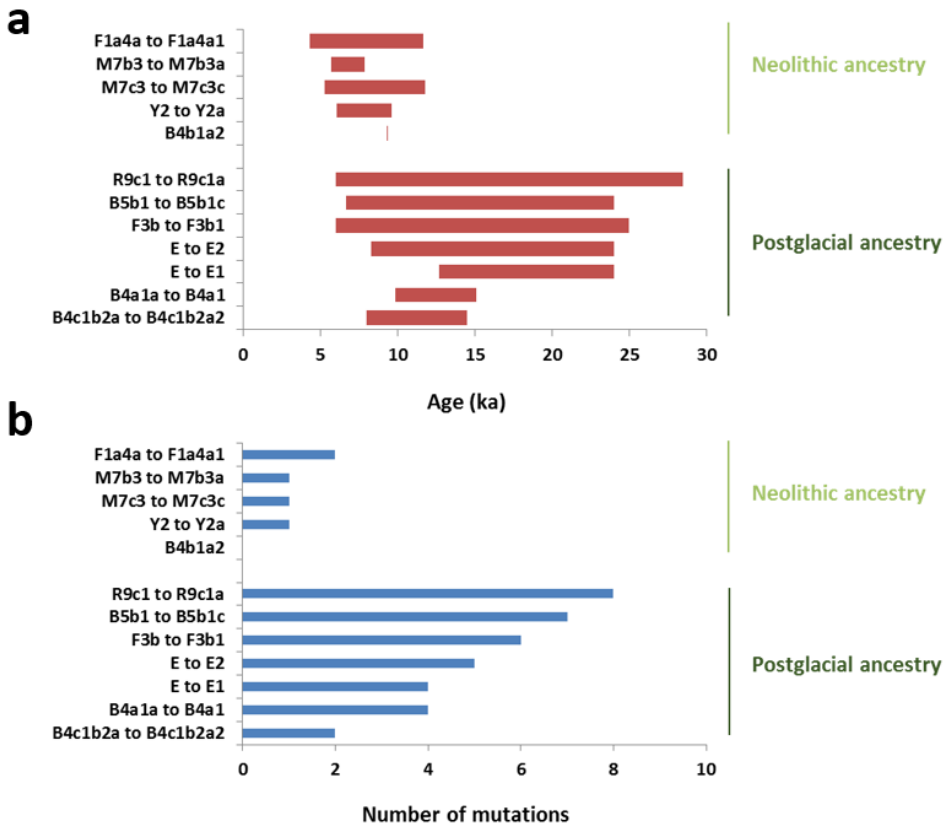


Figure S4. Phylogeographic patterns in ISEA. (a) ML ages of key mtDNA clades in ISEA and its ancestral node. (b) Number of mutations between key mtDNA clades in ISEA and its ancestral node.

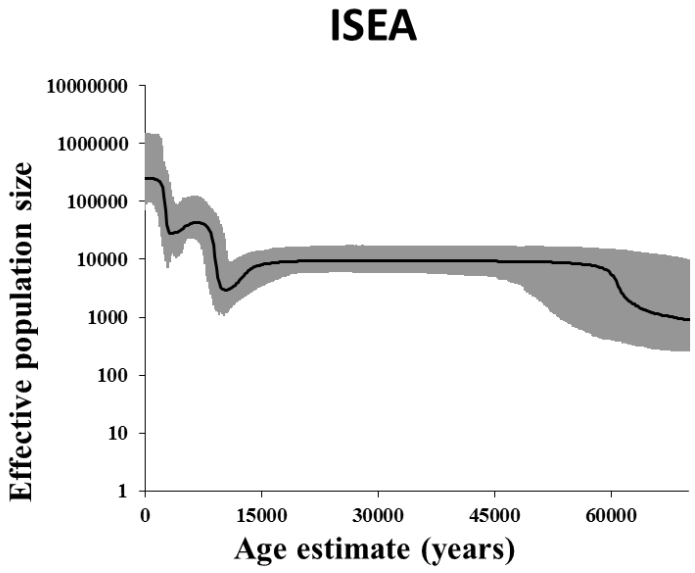


Figure S5. Bayesian skyline plots for ISEA, with the whole-mtDNA data set available, assuming a generation of 25 years. The black line represents the posterior effective population size through time, and the grey regions represents the 95% confidence interval.

Table S1. List of the 114 whole-mtDNA genomes sequenced and characterized in this study and corresponding geographic region.

Due to the large size of the table S1, the data is provided in digital format (Supplementary Tables – Paper II)

Table S2. List of the 829 published whole mitochondrial genomes used for the phylogeographic analysis with the corresponding origin and haplogroup affiliation.

Due to the large size of the table S2, the data is provided in digital format (Supplementary Tables – Paper II)

Table S3. Age estimates using rho (ρ) and ML for haplogroups B4b1, B4c1, B5b, D5, F1a4, F3, N9a, R9b, R9c and Y2, and its major subclades. Ages and 95% confidence intervals (CI) in thousands of years.

Due to the large size of the table S3, the data is provided in digital format (Supplementary Tables – Paper II)

Table S4. Entrance age estimates of the mtDNA lineages in this study in ISEA or Taiwan.

mtDNA lineages	Region	Age estimate (years)	95% confidence interval
N9a6a	ISEA	8,600	2,200 – 15,200
Y2a	ISEA	5,300	2,800 – 7,900
B4b1a2	ISEA	5,300	3,200 – 7,400
	Taiwan	6,700	4,300 – 9,200
B4c1b2a2	ISEA	7,600	4,900 – 10,400
	Taiwan	1,000	300 – 1,800
B5b1c	ISEA (Philippines)	8,900	6,500 – 11,300
R9b1a1a	ISEA	7,900	2,100 – 13,900
R9c1a	ISEA	5,900	3,300 – 8,600
	Taiwan	4,800	1,100 – 8,700
F1a4a1	ISEA	3,400	1,200 – 5,600
F3b1	Taiwan	5,800	700 – 11,100
D5b1c1a	ISEA	4,600	0 – 9,300

Table S5. Founder ages estimates for the main clades in Taiwan.

Founder clade	Age (years)	95% confidence interval
A5b1	0	–
B4a1a	5,900	2,500 –9,300
B4a2	10,100	5,300–15,000
B4b1a2	8,000	4,100–12,100
B5a2	14,100	6,500–22,000
C7a	0	–
D4a	0	–
D4i	0	–
D5b3	10,200	3,800–16,700
D6a2	0	–
E1a	7,100	3,600–10,600
E2b	4,200	1,400–7,100
F1a1d	4,600	300–8,900
F1a3	0	–
F1a3a	9,500	3,700–15,400
F1a4	3,500	0–7,100
F2	0	–
F3b1a	6,600	1,300–12,100
F4b	3,900	400–7,600
M7b1d3	5,600	1,100–10,300
M7b3a	7,700	4,500–10,900
M7c3a	5,700	2,400–9,100
M7c3c	7,400	4,500–10,400
M8a2	0	–
N9a10	8,300	2,300–14,600
R9b1a2	10,100	3,700–16,700
R9c1a	4,800	1,100–8,700
Y2a1a	4,900	1,800–8,000

Supplementary Note 1

Phylogeography of the mtDNA haplogroups targeted in this study

Here we provide, separately, the results of the phylogeographic reconstruction of all mtDNA haplogroups targeted in our study, as well the population expansion signals associated to each mitochondrial DNA lineage.

Haplogroup N9a

The overall phylogeographic pattern of mtDNA haplogroups N9 suggests an Eastern Asian origin ~50 ka (50.6 [37.1; 64.6] ka) (Table 1). N9 encompasses three basal branches: N9a, N9b and Y. The major subclade, N9a, dates to ~20 ka (20.0 [14.5; 25.5] ka) and is frequent across China, Japan and West ISEA (Fig. S1), suggesting an East Asian ancestry of this clade around the time of the LGM. N9a splits into six subclades, four with a strong Northeast Asian (i.e. Chinese and/or Japanese) focus (N9a1'3, N9a2'4'5, N9a8 and N9a9), and two with a probable MSEA/South Chinese ancestry (N9a6 and N9a10).

N9a6, dating to ~15 ka (14.8 [9.9; 19.8] ka), is by far the most frequent subclade of N9 in SEA. This subclade encompasses several basal branches with a strong MSEA ancestry, centred on Vietnam, and two major subclades, N9a6a and N9a6b, present in Austronesian-speaking populations. N9a6a, as shown previously (Hill et al. 2007), is frequent in Malaysia (N9a6a1) and the islands of Borneo and Sumatra (N9a6a2, N9a6a3 and N9a6a4) and, considering its estimated time of entrance in ISEA (8.6 [2.2; 15.2] ka) (Table S4), it probably expanded during the final flooding period (Pelejero et al. 1999).

In contrast with this pattern, subclade N9a10 arose within the region of South China and MSEA, in the same time frame (16.6 [11.0; 22.4] ka), but its major subclade, N9a10a, is mostly present in Taiwan and South China and absent in ISEA (apart from one individual so far in the Batanes Island, most probably related to Taiwan considering the close relationships between the Ivatan and the Yami people (Loo et al. 2011)). Considering the dispersal period of ~10-6 ka (10.0 [4.7; 15.4]; 6.3 [0.2; 12.5] ka) (Table 1) in the region, it seems plausible that N9a10 arrived in Taiwan from China with Neolithic rice

agriculturalists, together with M7b1d3, M7b3a and M7c3c (Soares et al. 2016)– but, similar to M7b1d3, there is no sign that it dispersed towards ISEA.

The BSP obtained for mtDNA haplogroup N9a shows two waves of population expansion (Fig. S2), at ~12 ka and ~7 ka (Table 2), the latter one correlating well with the expansion into ISEA and Malaysia, as given by the age of N9a6a. Given the overall phylogenetic and phylogeographic pattern and the BSP population expansions, mtDNA haplogroup N9a appears to have an East Asian origin with some of its lineages spreading to the Sunda continent during the postglacial period, thus constituting a strong candidate for a postglacial mtDNA signal for migrations from MSEA.

Haplogroup Y

A second branch of N9, haplogroup Y, dates to almost 30 ka (28.0 [16.1; 40.5] ka) (Table 1). This splits into two major subclades, Y1, which shows a clear North Asian ancestry, probably within South Siberia at the end of Last Glacial period, and a second younger subclade Y2, dating to ~10 ka (9.6 [5.0; 14.4] ka), which is mostly found in Taiwan and ISEA, at an overall frequency of ~18%. Y2a, dating to ~6 ka (6.1 [3.2; 9.1] ka), is the most frequent branch of Y2, and the only one observed in Austronesian-speaking populations. The age interval between Y2 (9.6 ka) and Y2a (6 ka) suggests a migration from mainland to insular locations in the time frame of the hypothetical settlement of Taiwan by rice-agriculturalists from South China. Y2a is frequent in Taiwan and ISEA, where its founder age suggested an entrance in ISEA around ~5 ka (5.3 [2.8; 7.9] ka). Y2a encompasses a “star-like” branch, Y2a1a that is mainly found in ISEA which, given its age (4.1 [2.3; 5.8] ka), seems likely to have accompanied the Neolithic Austronesian dispersal from Taiwan into ISEA. The BSP also displays a population increment after 5 ka. Y2, mostly Y2a1, thereby represents, along with M7c3c (Soares et al. 2016), a strong candidate for an OOT mtDNA marker.

Haplogroup B4b1

Haplogroup B4 (Hill et al. 2007) is subdivided into three main subclades, B4a, B4b'd and B4c, and dates to ~35 ka (Derenko et al. 2012). Subclade B4b splits in two major subclades, the Amerindian-specific branch B2 (Bandelt et al. 2003; Eshleman et al. 2003) and B4b1, dating to ~25 ka (25.1 [17.0; 33.6] ka) (Table 1), which displays a wide geographic distribution from Northeast Asia to SEA (Fig. S1).

We previously identified B4b1 as a candidate genetic marker of the pottery-making rice-farming OOT dispersal into ISEA (Soares et al. 2016). The reconstructed whole-mtDNA genome phylogeography shows that within B4b1, B4b1a is by far the most frequent subclade in SEA, found in both MSEA and ISEA. This subclade is also found in Northern Asia, in the form of B4b1a1 and B4b1a3.

B4b1a2, dating to ~9 ka (9.3 [6.8; 11.8] ka), is the major subclade of B4b1a, and is the only one that is present in Taiwan and ISEA. A founder age from ISEA into Taiwan indicates an entrance ~7 ka (6.7 [4.3; 9.1] ka), matching the hypothetical rice-agriculturalist migration from China to Taiwan. Given the overall phylogeographic patterns of this clade, with a founder age of ~5 ka (5.3 [3.2; 7.4] ka) in ISEA, it seems possible that B4b1a2 could have dispersed from South China into Taiwan and later to ISEA in a similar fashion to M7c3c, following the OOT model. Another similarity with M7c3c is that the clade expanded to Micronesia and northwest Polynesia and is not detected along the North Coast of New Guinea or the Bismarck Archipelago. The increment time shown by the BSP (Table 2) within the Neolithic time frame also supports B4b1a2 as a genetic signal for the Neolithic dispersal from Taiwan to ISEA.

Haplogroup B4c1

mtDNA haplogroup B4c has a pre-LGM Northern Asian origin, followed by a later distribution of its major subclade B4c1 and minor subclade B4c2 (Derenko et al. 2012). B4c1 is broadly frequent throughout East Asia and SEA (Fig. S1). It splits into two further subclades: a minor Japanese offshoot, B4c1c, and a second major subclade, B4c1a'b, incorporating B4c1a, displaying a clear Northeast Asian ancestry centred on Japan, and B4c1b which, by contrast, is prominent throughout Malays, Filipinos and aboriginal Taiwanese. Within B4c1b, the subclade B4c1b2 – more specifically the branch B4c1b2a2, dating to ~8 ka (8.0[5.6; 10.5] ka) (Table 1) – is by far the most frequent lineage in Taiwan and ISEA, suggesting an ancestry within those regions.

Soares et al. (2016), based on mtDNA HVS-I data, suggested that B4c1 could represent a genetic marker for the OOT model. At the whole-mtDNA genome level, only one subclade of B4c1b2a2 is found in Taiwan, whereas several exist in ISEA, including the branch that appears in Taiwan (detected in the Philippines). This evidence points to an origin in ISEA and a northwards migration. Using an exploratory founder analysis, considering ISEA and Taiwan as hypothetical source and sink populations and *vice versa,* we estimated a founder age of ~1 ka (1.0 [0.3; 1.7] ka) in Taiwan and 7.6 [4.9; 10.4] ka in ISEA,

emphasizing that an origin in Taiwan is very unlikely. The increment ~6 ka in the BSP (Table 2), as for B4b1 above, does not clearly distinguish between an OOT or a postglacial expansion. The age of B4c1b2a at 14.5 [6.4; 23.1] ka indicates that the clade entered ISEA between 15 ka and 8 ka, a pattern similar to the more common B4a1a1 (Soares et al. 2011).

In fact, the B4c1b2a2a subclade (defined here) is the only Taiwanese branch of B4c1, and is found only in the Yami. These people, also known as Tao, are native to the small outlying Orchid Island in Taiwan and are distinct from other Taiwanese aboriginal groups as the only non-Formosan Austronesian speakers among Taiwanese aborigines. The languages of Yami belong to the Batanic sub-branch of the Western Malayo-Polynesian branch, which comprises all of the Austronesian languages spoken outside Taiwan(Blust 2009), suggesting a recent migration from the Philippines. This hypothesis is supported by the recent founder age ~1 ka as calculated above. Given the overall pattern, B4c1b2a2 appears to constitute a genetic signature of the reticulated network of cultural/linguistic relationships between Orchid Island and Philippines previously described by Ross (Ross 2005). Considering that B4c1b2a2 is not defined by any HVS-I variants, the diversity detected outside ISEA was probably part of B4c1b2a* which allowed the clade to be considered a founder into ISEA in the HVS-I founder analysis (Soares et al. 2016) under the founder analysis criteria for derived clades in the source (Richards et al. 2000). This reinforces the need to study phylogeographic patterns at the whole-mtDNA level for optimal results.

Haplogroup B5b

The mtDNA haplogroup B5b is the other Asian haplogroup B lineage that we are focusing on in our study. B5b reaches up to ~8–9% in Malaysia and southeast Indonesia, but it is considerably less frequent in the neighbouring regions (Fig. S1). B5b dates to ~30 ka (29.8 [20.7; 39.3] ka) (Table 1), and the major subclade, B5b1, diverged ~24 ka (23.9 [13.9; 34.3] ka) somewhere within East Asia. B5b1 splits into two main subclades: B5b1a is an entirely Japanese clade that dates to ~19 ka (19.0 [8.8; 29.8] ka) and B5b1c is found only in Austronesian-speaking populations, with a founder estimate age in the Philippines ~9 ka (8.8 [6.5; 11.3] ka). The time gap between this Holocene subclade and its Late Pleistocene ancestor (B5b1) suggests extensive genetic drift during this period, most likely due to the sea-level rises that resulted from the global warming at the end of the

Pleistocene (Pelejero et al. 1999). The BSPs also show two-stepped population growth in the Late Pleistocene and Holocene (Table 2).

Given the overall pattern, B5b seems to have had a South Chinese origin in the Late Pleistocene and to have spread widely over central/eastern Asia since then, similar to B4c1b2a2 (above) and B4a1a (Soares et al. 2011),and its arrival in ISEA was prompted by climate change, rather than driven by the Neolithic (from MSEA in this case, not Taiwan) as inferred from the HVS-I founder analysis performed previously (Soares et al. 2016). Again, the fact that B5b1c (or even B5b1) is not defined by any HVS-I mutation led the clade as a whole to be considered a founder from Asia, resulting in very imprecise age estimates.

Haplogroup R9b

mtDNA haplogroup R9 encompasses three basal branches, R9b, R9c and F, all with an East Asian origin in the Late Pleistocene (Hill et al. 2006; Peng et al. 2010). R9b is frequent throughout East Asia, reaching ~20% in South China. The presence of this clade gradually decreases throughout continental and insular Southeast Asia, and it is completely absent in the Philippines (Fig. S1). R9b was identified as a possible marker for postglacial expansions by Hill et al. (2006). Here we reanalysed the R9b phylogeography in light of the whole-mtDNA sequences, as well as our re-evaluation of the molecular clock (Soares et al. 2009).

R9b dates to ~39 ka (38.7 [23.9; 54.3]) ka (Table 1), and splits into two main branches: R9b1, with a clear, ancient MSEA ancestry, and a more recent subclade, R9b2, dating to just ~6 ka (5.7 [1.3; 10.2] ka), only found in Malays, Vietnamese and Thai. This implies an overall origin of R9b in MSEA. R9b1a, the major clade of R9b, dating to ~18 ka (18.6 [10.8; 26.7] ka), splits into three subclades, R9b1a1, R9b1a2 and R9b1a3. The phylogeographic distribution of R9b1a1 suggests an expansion within the SEA ~11 ka (11.6 [6.0; 17.3] ka), with its subclade, R9b1a1a, showing a postglacial founder estimate age in insular Southeast Asia of ~8 ka (7.9 [2.1; 13.9] ka).

This overall pattern, allied to the population expansion detected in the Mid-Holocene (Fig. S2), supports the view of R9b as a genetic marker for postglacial expansions from MSEA, as suggested before (Hill et al. 2006). The mtDNA haplogroup R9b – similarly to N9a – has an East Asian origin in the Late Pleistocene, and later with the climatic improvement

dispersed to MSEA and southwards towards Malaysia and ISEA at the time of sea-level rises.

Haplogroup R9c

R9c, which is mostly found on the eastern side of ISEA (Fig. S1), has only one major subclade, R9c1, dating to ~28 ka (28.4 [17.2-40.3] ka) (Table 1). This subclade splits into three main branches, all of them with a tree structure and distribution focused on ISEA. The main branch R9c1a, dating to ~6 ka (5.9 [3.7; 8.3] ka), is largely restricted to Austronesian-speaking populations, excepting one South Chinese sample. Following previous inferences (Soares et al. 2016), R9c1a appears to have entered ISEA at the end of the postglacial dispersals (~5.9 [3.3; 8.6] ka), later reaching Taiwan (~4.8 [1.1; 8.7] ka). This was mainly inferred from a distribution centred on the Sunda shelf and the complete lack of HVS-I diversity in Taiwan. However at the whole-mtDNA level the picture becomes less clear-cut. There is a single South Chinese sample that could represent an accidental or indicate an origin in South China/Taiwan. Also, the age estimate of the clades at ~6 ka is ambiguous and slightly older than the hypothetical OOT migration, but it is also more recent than the postglacial expansions. The large age distance between R9c1a and its ancestor, R9c1 (nearly 30 ka (28.5 [17.2-40.3] ka)) indicate that this subclade could have been within ISEA or Taiwan since the Late Pleistocene, a pattern mainly observed in clades present in ISEA that went through extensive genetic drift during the flood episodes (as B4a1a1, E, B4c1b2a2, and B5b1). The recent population expansion (~2 to 5 ka) (Table 2) detected by the BSP could indicate a more recent autochthonous Southeast Asian expansion.

Overall, R9c does not fit the OOT phylogeographic parameters established by Soares et al. (2016) and shows ambiguous results; but the overall pattern suggests an ISEA origin rather than an OOT origin.

Haplogroup F1a4

The mtDNA haplogroup F1a4 is extremely rare, found at low frequency only in South China and Austronesian-speaking populations. F1a4 dates to 16.2 [7.0; 26.0] ka (Table 1) and includes a major subclade, F1a4a, with a clear Chinese origin ~12 ka (11.7 [3.0; 20.9] ka) (Table 1). This clade encompasses a star-like cluster of several Taiwanese ancestors, referred as F1a4a1, dating to just above 4 ka (4.3 [1.8; 6.8] ka), and a descendent

subclade, F1a4a1a, found only in ISEA, Malaysia and Micronesia, dating just over 3 ka (3.3 [1.3; 5.3] ka). The time gap between the emergence of F1a4a and this F1a4a1 suggests a time of arrival in Taiwan fitting well the time frame for the arrival of rice-agriculturalists in the OOT model. An entry into ISEA between the age of F1a4a1 in Taiwan (4.3 ka) and F1a4a1a at 3.3 ka also fits the Austronesian migration in the OOT model. The BSP plot of F1a4 shows population increment within the last 4 ka (Table 2), which fits the arrival/dispersal time of F1a4a1 in ISEA (~3.4 [1.2; 5.6] ka). Given this overall pattern, F1a4 could have been carried into Taiwan from South China by rice-agriculturists, and later with the OOT migration into ISEA along with mtDNA lineages B4b1a2, Y2a1 and M7c3c (Soares et al. 2016). As with M7c3c and B4b1a2, the presence of this clade in the Pacific is evident only in Micronesia (suggesting an arrival there directly from ISEA).

Haplogroup F3

The mtDNA haplogroup F3 dates to ~32 ka (31.7 [21.5; 42.3] ka) (Table 1) and is fairly common throughout East and Southeast Asia. There are two major basal subclades, F3a and F3b, both with similar Late Pleistocene ages (26.7 [16.5; 37.1] ka and 25.2 [15.4; 35.4] ka, respectively) (Table 1), but with clearly different ancestries. F3a is mostly present in MSEA, such as Vietnam, Laos, Malaysia and southern China, suggesting that this clade has a MSEA ancestry. The daughter clade, F3a1, dates to ~16 ka (16.6 [9.0; 24.5] ka) and, similarly to its ancestral clade, displays a MSEA origin centred on Vietnam and Laos.

The sister clade F3b is divided into two subclades. One, F3b2, is rare and was detected only in South China, while F3b1, dating to ~12 ka (12.4 [5.2; 20.0] ka), is by far the more common subclade of F3b and is largely restricted to Austronesian-speaking populations in ISEA and Taiwan. Within F3b1, F3b1b is restricted to ISEA while F3b1a is found in ISEA and Taiwan, strongly suggesting an origin in ISEA and a migration into Taiwan. A founder age into Taiwan (5.8 [0.7; 11.1] ka) (Table S4), is concordant with the hypothesis that this clade accompanied postglacial dispersed from ISEA towards Taiwan, again most probably as a result of sea-level rises. Overall, indeed, the age and distribution of haplogroup F3 shows many similarities with haplogroup E (Soares et al. 2008). It likewise emerged in ancient Sundaland over 30 ka, but probably further to the west, within what is now MSEA. Two subclades within haplogroup F3b1 show traces of expansion in the last 8 ka in ISEA, with one reaching Taiwan. The BSP for mtDNA haplogroup F3 shows two population

expansion periods, the first between ~5–10 ka and the second within the last 4 ka (Table 2). Given the phylogeographic and phylogenetic patterns of F3 overall, it seems likely that this clade dispersed more than once within the Sunda region over the last ~16 ka.

Haplogroup D5

The mtDNA haplogroup D5 dates to just over ~30 ka (33.3 [24.6; 42.2] ka), and is widely distributed throughout East and Southeast Asia. There are two basal branches, D5a'b and D5c. The latter further separates into D5c1 and D5c2, both with a probable North/Northeast Asian origin. D5a'b separates into two major subclades: D5a, which is widely dispersed throughout East and Northeast Asia, and D5b, which is extremely frequent in Taiwan and less frequent in Southeast Asia. This clade splits into two subclades, D5b1 and D5b3 (a newly defined branch). D5b3 dates to ~11 ka (10.9 [5.6; 16.4] ka) (Table 1) and is largely restricted to Chinese and Taiwanese populations, and virtually absent in ISEA. Given the existence of several Taiwanese branches dating to less than 4 ka, it seems likely that D5b3 moved between 10 ka and 3 ka, suggesting that it could have arrived in Taiwan with the Neolithic rice-farmers from South China. However, it did not follow the Austronesian movement OOT, resembling in thsi respect the patterns of mtDNA haplogroups N9a10a and M7b1d3.

Within D5b1, subclade D5b1c1, dating to ~9 ka (9.1 [4.0; 14.4] ka) is the only D5 subclade to disperse to insular Southeast Asia. This subclade includes a cluster with ancestry in Taiwan, D5b1c1a, dating ~6 ka (6.0 [0; 13.8] ka), restricted to Austronesian-speaking populations. Although the tree might seem to imply a deeper ancestry in ISEA than in Taiwan, this is caused by a single HVS-I variant, 16092, that is mildly fast and could represent homoplasy. A founder age into ISEA is ~4.6 [0; 9.3] ka, again suggesting a Neolithic OOT marker clade. The population increase between ~13 ka till ~3.5 ka, with a peak at ~7.7 ka (Table 2), mostly shows a signal of early population expansion within South China. In contrast to other clades described above, the Austronesian component in the BSP is somewhat low which does not make any hypothetical OOT expansion important in the overall BSP against postglacial expansions in continental Asia. Although it is a probable OOT marker its presence in ISEA is low, at comparable levels to another OOT candidate, M7b3 (Soares et al. 2016).

References

Bandelt HJ, Herrnstadt C, Yao YG, Kong QP, Kivisild T, Rengo C, Scozzari R, Richards M, Villems R, Macaulay V, Howell N, Torroni A, Zhang YP (2003) Identification of Native American founder mtDNAs through the analysis of complete mtDNA sequences: some caveats. Ann Hum Genet 67:512-524.

Blust R (2009) The austronesian languages. Pacific Linguistics, Canberra, Australia. 824 pp.

Derenko M, Malyarchuk B, Denisova G, Perkova M, Rogalla U, Grzybowski T, Khusnutdinova E, Dambueva I, Zakharov I (2012) Complete mitochondrial DNA analysis of eastern Eurasian haplogroups rarely found in populations of northern Asia and eastern Europe. PLoS One 7:e32179. doi:10.1371/journal.pone.0032179

Eshleman JA, Malhi RS, Smith DG (2003) Mitochondrial DNA studies of Native Americans: conceptions and misconceptions of the population prehistory of the Americas. Evol Anthropol 12:7-18.

Hill C, Soares P, Mormina M, Macaulay V, Clarke D, Blumbach PB, Vizuete-Forster M, Forster P, Bulbeck D, Oppenheimer S, Richards M (2007) A mitochondrial stratigraphy for Island Southeast Asia. Am J Hum Genet 80:29-43. doi:10.1086/510412

Hill C, Soares P, Mormina M, Macaulay V, Meehan W, Blackburn J, Clarke D, Raja JM, Ismail P, Bulbeck D, Oppenheimer S, Richards M (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. Mol Biol Evol 23:2480-2491. doi:10.1093/molbev/msl124

Loo J-H, Trejaut JA, Yen J-C, Chen Z-S, Lee C-L, Lin M (2011) Genetic affinities between the Yami tribe people of Orchid Island and the Philippine Islanders of the Batanes archipelago. BMC Genet 12:21.

Pelejero C, Kienast M, Wang L, Grimalt JO (1999) The flooding of Sundaland during the last deglaciation: Imprints in hemipelagic sediments from the southern South China Sea. Earth and Planetary Science Letters 171:661-671.

Peng M-S, Quang HH, Dang KP, Trieu AV, Wang H-W, Yao Y-G, Kong Q-P, Zhang Y-P (2010) Tracing the Austronesian footprint in Mainland Southeast Asia: a perspective from mitochondrial DNA. Mol Biol Evol:msq131.

Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. Am J Hum Genet 67:1251-1276.

Ross M (2005) The Batanic languages in relation to the early history of the Malayo-Polynesian subgroup of Austronesian. Journal of Austronesian Studies 1:1-24.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, Salas A, Oppenheimer S, Macaulay V, Richards MB (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet 84:740-759.

Soares P, Rito T, Trejaut J, Mormina M, Hill C, Tinkler-Hundal E, Braid M, Clarke DJ, Loo J-H, Thomson N, Denham T, Donohue M, Macaulay V, Lin M, Oppenheimer S, Richards MB (2011) Ancient voyaging and polynesian origins. Am J Hum Genet 88:239-247. doi:10.1016/j.ajhg.2011.01.009

Soares P, Trejaut JA, Loo JH, Hill C, Mormina M, Lee CL, Chen YM, Hudjashov G, Forster P, Macaulay V, Bulbeck D, Oppenheimer S, Lin M, Richards MB (2008) Climate change and postglacial human dispersals in Southeast Asia. Mol Biol Evol 25:1209-1218. doi:10.1093/molbev/msn068

Soares P, Trejaut JA, Rito T, Cavadas B, Hill C, Eng KK, Mormina M, Brandão A, Fraser RM, Wang T-Y, Loo J-H, Snell C, Ko T-M, Amorim A, Pala M, Macaulay V, Bulbeck D, Wilson JF, Gusmão L, Pereira L, Oppenheimer S, Lin M, Richards MB (2016) Resolving the ancestry of Austronesian-speaking populations. Hum Genet 135:309-326. doi:10.1007/s00439-015-1620-z

APPENDIX C – SUPPLEMENTARY INFORMATION OF PAPER III

The genome-wide landscape of Island Southeast Asia

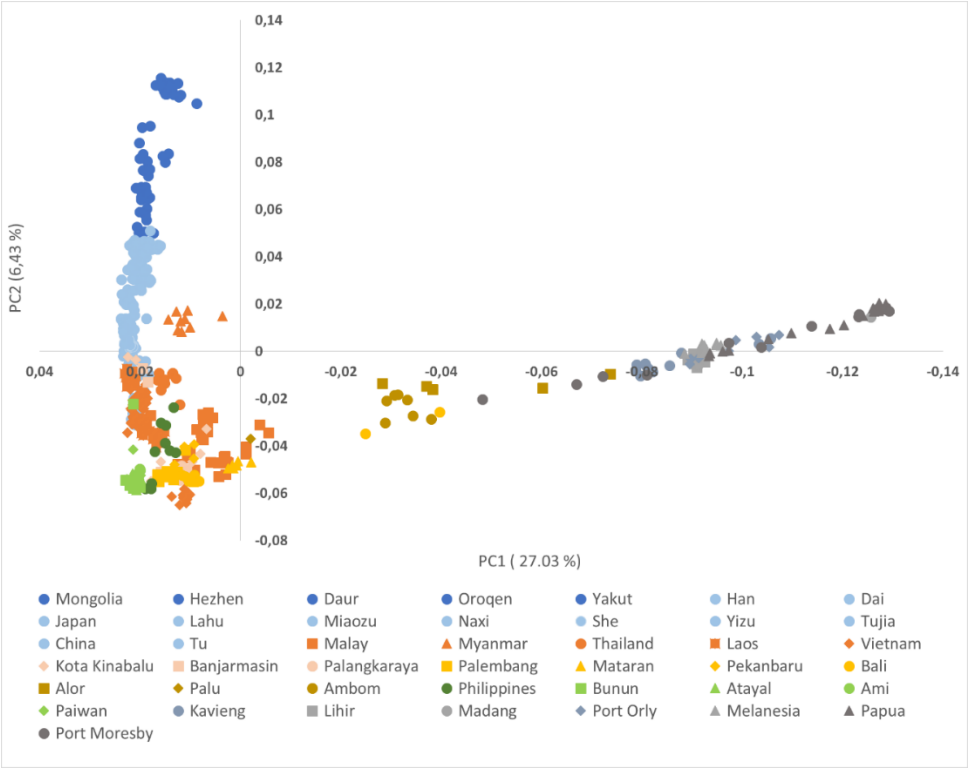


Figure S1. PCA plots for East and Southeast populations.

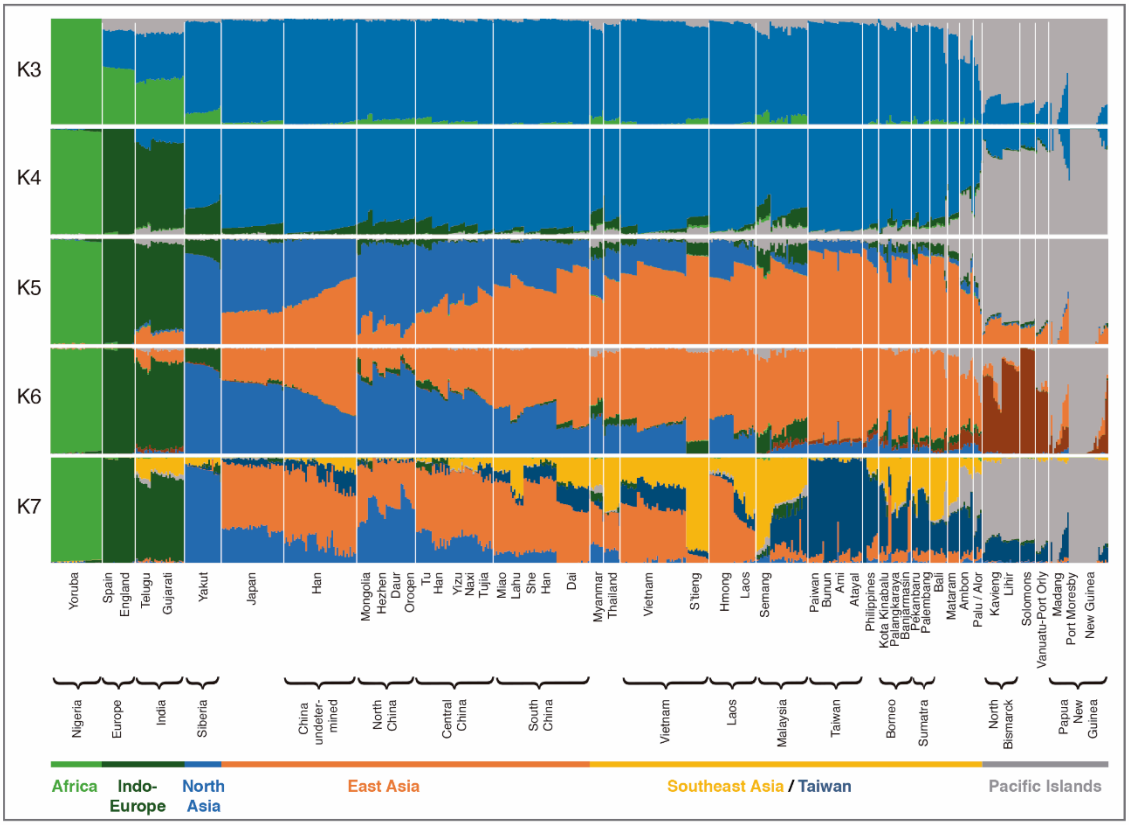


Figure S2. sNMF analysis (K=2 to K=7) of Asian and neighbouring populations and subpopulations. Each individual from populations is represent in the *x*-axis, as a vertical stacked column of color-coded admixture proportions of the putative ancestral populations.

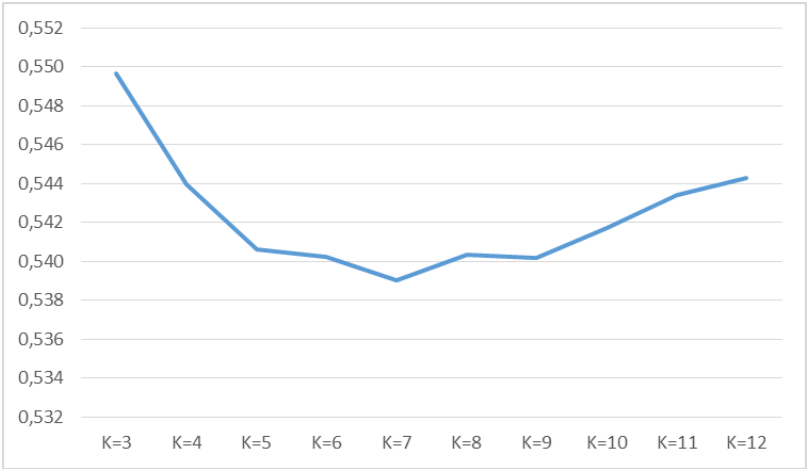


Figure S3. Cross-validation for ADMIXTURE analysis. Ks between 3 and 12.

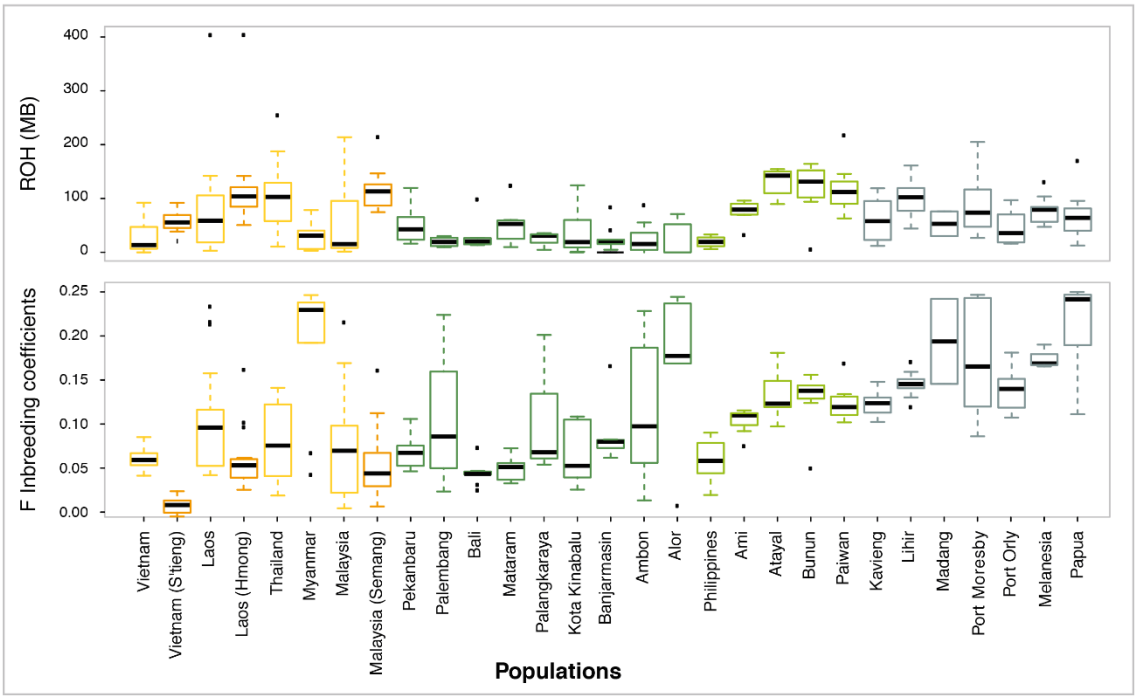


Figure S4: Box plot of the total ROH (Mb) and inbreeding factor (F) in the Southeast Asian populations. The bottom and the top are the first and third quartiles, whereas the line inside the box is the median. The whiskers are the maximum and the minimum and of all the data; and dots represent outliers not included in the whiskers.

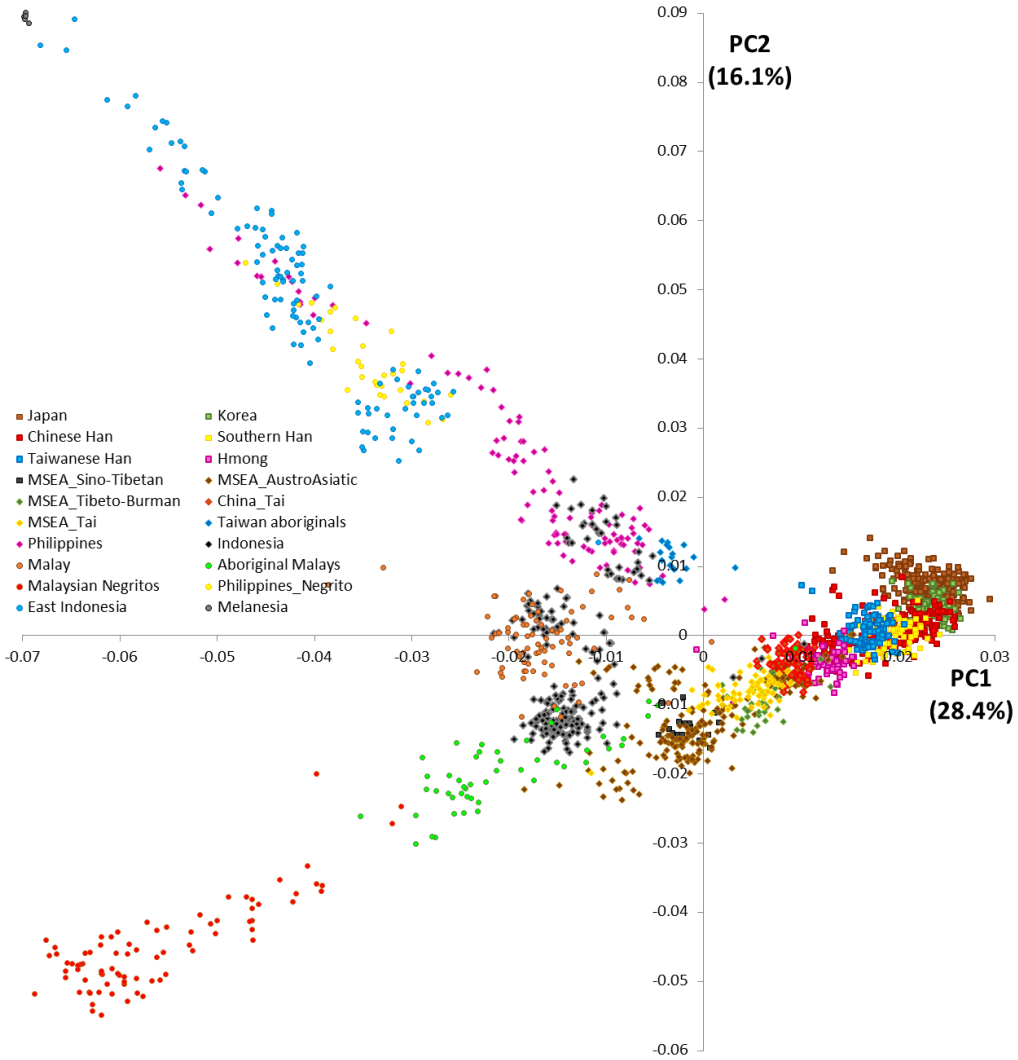


Figure S5. PCA plots for East and Southeast populations using the expanded Pan-Asian SNP Consortium (EPASC) dataset.

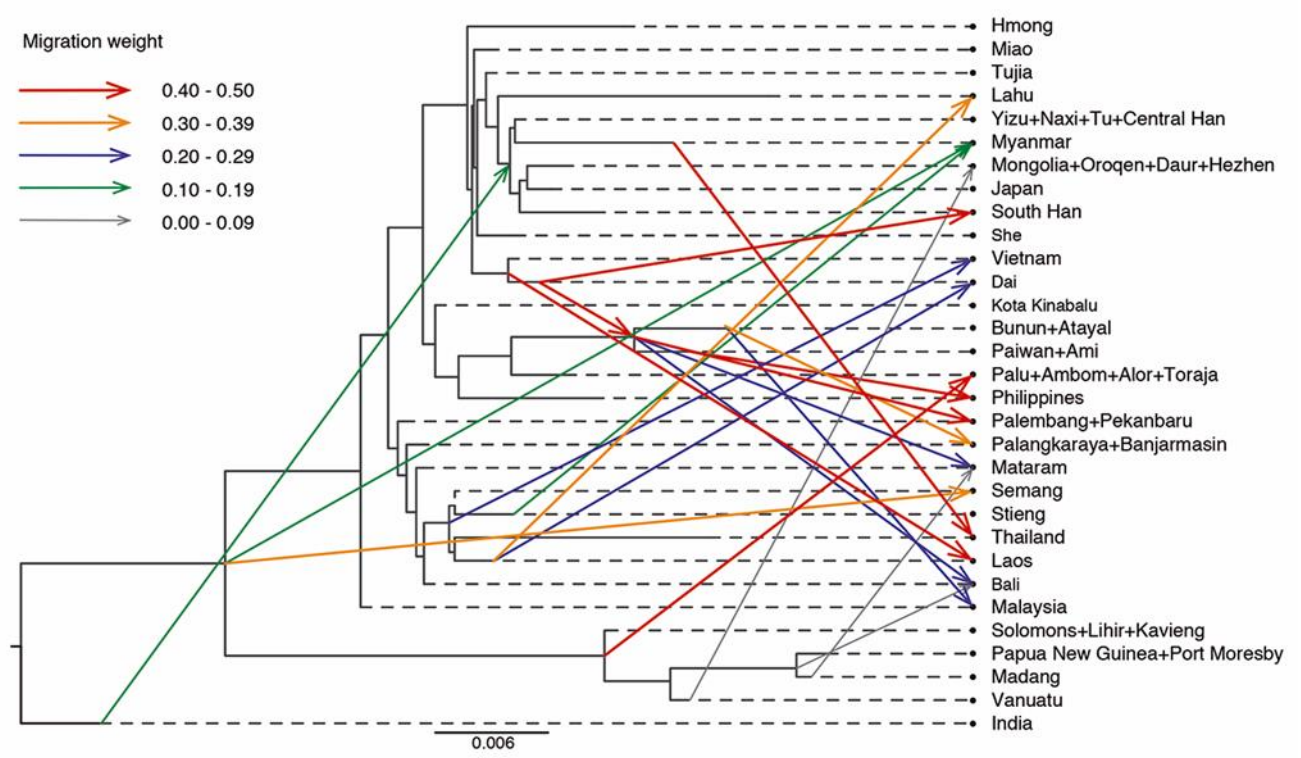


Figure S6. Maximum likelihood population tree and admixture events inferred by TreeMix. The tree that best fit the dataset has six inferred migration edges, which explains 99.7% of genetic variation of the populations. The spectrum of colour of the migration arrows indicates different migration weights. The branch lengths are proportional to the amount of genetic drift that has occurred on populations.

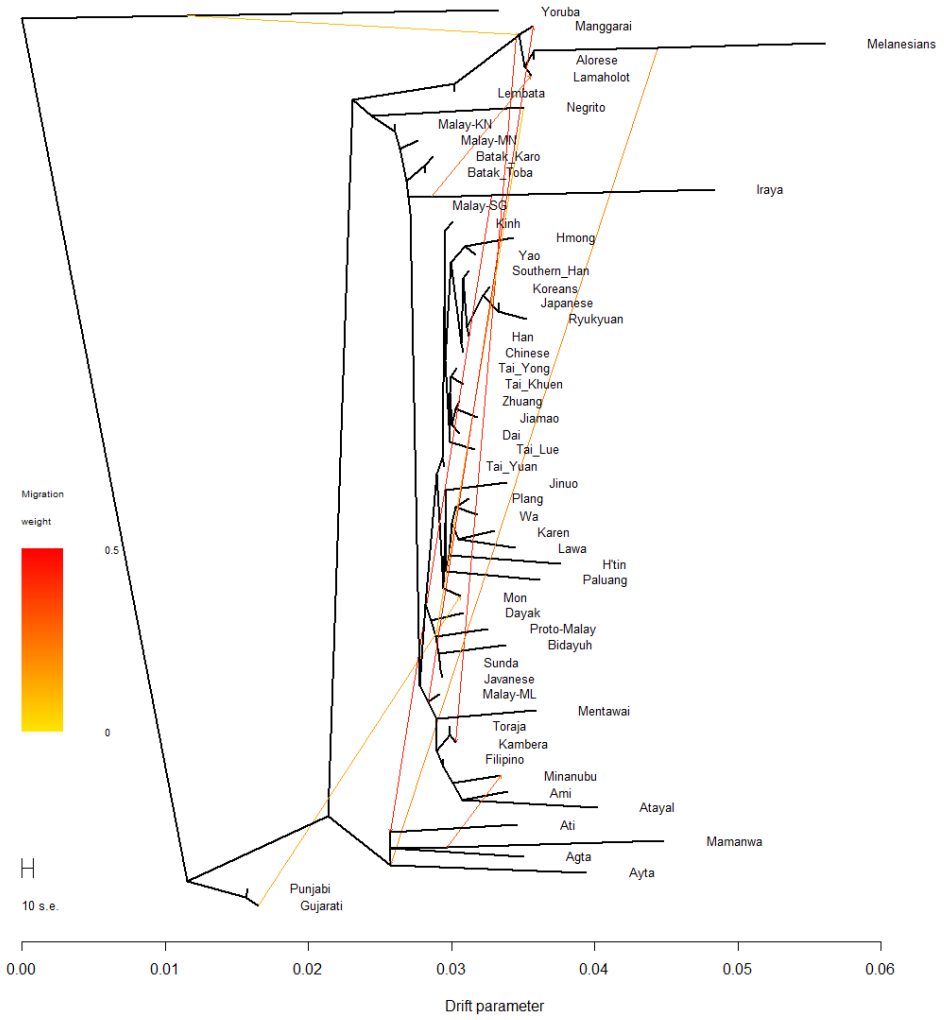


Figure S7. Maximum likelihood population tree and admixture events inferred by TreeMix for the expanded Pan-Asian SNP Consortium (EPASC) dataset. The spectrum of colour of the migration arrows indicates different migration weights.

Table S1. Characterization of the samples used included in this study.

Geographic division	Region	Population	# samples	Dataset Reference
Africa	Yoruba	Yoruba	21	HGDP-CEPH
			10	1000 Genomes Project
South Asia	India	Indian Gujarati	20	1000 Genomes Project
		Indian Telugu	10	1000 Genomes Project
Northeast Asia	China	Daur	9	HGDP-CEPH
	China	Hezhen	8	HGDP-CEPH
	China	Mongolia	10	HGDP-CEPH
	China	Oroqen	9	HGDP-CEPH
	China	Tu	10	HGDP-CEPH
	China	Yakut	23	HGDP-CEPH
East Asia	China	Chinese Dai	10	1000 Genomes Project
			10	HGDP-CEPH
		Chinese Han	20	1000 Genomes Project
			52	HGDP-CEPH
		Miaozu	10	HGDP-CEPH
		Naxi	8	HGDP-CEPH
		She	10	HGDP-CEPH
		Tujia	10	HGDP-CEPH
		Yizu	10	HGDP-CEPH
	Japan	Japan	28	HGDP-CEPH
			10	1000 Genomes Project
Taiwan	Taiwan	Ami	7	This study
		Atayal	7	This study
		Bunun	9	This study
		Paiwan	10	This study
Island Southeast Asia	East Indonesia	Alor	5	This study
	East Indonesia	Ambon	7	This study
	East Indonesia	Palu	1	This study
	East Indonesia	Toraja	1	This study
	Philippines	Philippines	10	This study
	West Indonesia	Bali	10	This study
	West Indonesia	Mataram	7	This study
	West Indonesia	Palembang	8	This study
	West Indonesia	Pekanbaru	4	This study
	West Indonesia (Borneo)	Banjarmasin	9	This study
	West Indonesia (Borneo)	Kota Kinabalu	8	This study
	West Indonesia (Borneo)	Palangkarya	3	This study
Mainland	Myanmar	Myanmar	9	This study

Southeast Asia	Laos	Laos	28	This study
	Malaysia	Malaysia	32	This study
	Thailand	Thailand	10	This study
	Vietnam	Kinh	15	This study
		Kinh	10	1000 Genomes Project
		S'tieng	14	This study
		Tay Nung	15	This study
Melanesia	Melanesia	Melanesian	10	HGDP-CEPH
	New Guinea	Kavieng	12	This study
		Lihir	11	This study
		Madang	2	This study
		Papua	3	HGDP-CEPH
		Papua	21	This study
		Port Moresly	10	This study
	Vanuatu	Port Orly	8	This study

Table S2. Characterization of the samples of expanded Pan-Asian SNP Consortium (EPASC) dataset.

Geographic division	Region	Population	# samples	Dataset Reference
Africa	Yoruba	Yoruba	60	1000 Genomes Project
South Asia	India	Gujarati	103	1000 Genomes Project
		Punjabi	96	1000 Genomes Project
East Asia	China	Chinese	110	Pan-Asian SNP
		Dai	93	1000 Genomes Project
		Han	78	1000 Genomes Project
			51	Pan-Asian SNP
		Southern Han	105	1000 Genomes Project
		Hmong	26	Pan-Asian SNP
		Jiamao	31	Pan-Asian SNP
		Jinuo	29	Pan-Asian SNP
		Wa	56	Pan-Asian SNP
		Zhuang	26	Pan-Asian SNP
	Japan	Japanese	71	Pan-Asian SNP
			104	1000 Genomes Project
		Ryukyuan	49	Pan-Asian SNP

	Korea	Korean	90	Pan-Asian SNP
Taiwan	Taiwan	Ami	10	Pan-Asian SNP
		Atayal	10	Pan-Asian SNP
Island Southeast Asia	Indonesia	Alorese	19	Pan-Asian SNP
		Dayak	12	Pan-Asian SNP
		Javanese	53	Pan-Asian SNP
		Batak Karo	17	Pan-Asian SNP
		Kambera	20	Pan-Asian SNP
		Lamaholot	20	Pan-Asian SNP
		Lembata	19	Pan-Asian SNP
		Malay	12	Pan-Asian SNP
		Mentawai	15	Pan-Asian SNP
		Manggarai	36	Pan-Asian SNP
		Sunda	25	Pan-Asian SNP
		Batak Toba	20	Pan-Asian SNP
		Toraja	20	Pan-Asian SNP
	Philippines	Agta	8	Pan-Asian SNP
		Ayta	8	Pan-Asian SNP
		Ati	23	Pan-Asian SNP
		Iraya	9	Pan-Asian SNP
		Minanubu	18	Pan-Asian SNP
		Mamanwa	19	Pan-Asian SNP
		Filipino	59	Pan-Asian SNP
Mainland Southeast Asia	Malaysia	Bidayuh	50	Pan-Asian SNP
		Negrito	80	Pan-Asian SNP
		Malay	38	Pan-Asian SNP
		Proto-Malay	49	Pan-Asian SNP
	Singapore	Malay	30	Pan-Asian SNP
	Thailand	Hmong	20	Pan-Asian SNP
		Karen	20	Pan-Asian SNP
		Lawa	19	Pan-Asian SNP
		Mon	19	Pan-Asian SNP
		Paluang	18	Pan-Asian SNP
		Plang	18	Pan-Asian SNP
		Tai Khuen	18	Pan-Asian SNP
		Tai Lue	20	Pan-Asian SNP
		H'tin	18	Pan-Asian SNP
		Tai Yuan	20	Pan-Asian SNP
		Tai Yong	18	Pan-Asian SNP
		Yao	19	Pan-Asian SNP
	Vietnam	Kinh	99	1000 Genomes Project
Melanesia	Melanesia	Melanesian	5	Pan-Asian SNP